REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Sept. 20, 1997	Final Report 8/1/93 - 7/31/97

4. TITLE AND SUBTITLE
Automatic Target Recognition and Indexing by Non-Orthogonal Image Expension and Data-Dependent Normalization with Implementation

5. FUNDING NUMBERS
C
N00014-93-1-1088

6. AUTHORS
Professor J. Ben-Arie, Principal Investigator, University of Illinois at Chicago
Professor G. Atkin, Principal Investigator, Illinois Institute of Technology

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Illinois Institute of Technology, 3300 S. Federal Street, Chicago, IL 60616
University of Illinois at Chicago, (M/C 672) 1737 W. Polk Street, chicago, IL 60612

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Department of the Navy, Office of Naval Research
Chicago Region Office, Federal Buildign Room 208
536 South Clark Street, Chicago, IL 60605-1588

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
References [1] to [7] in Final Report.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

Per Thomas McKenna ONR/Code 342 1/7/98

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This research is concerned with the development of a neural system for robust projective-invariant recognition of multiple targets which may be partially occluded in a cluttered background based on single gray-level images. For this purpose we have developed in the research a new method for affine-invariant iconic representation and recognition of targets using a novel set of Gabor/Fourier kernels with multi-dimensional indexing in the frequency domain. An affine-invariant representation of local image patches is extracted in the form of spectral signatures, by directly convolving the image with our novel configuration of these kernels.

We achieved 100% correct recognition rates with a model library of 26 models over a wide range of viewing poses and distances (360 ° of rotation and tilt and 82 ° of slant and 4 octaves of scale). The system also maintains its 100% recognition rate in high levels of noise/clutter (up to -17 dB) and to resolution degradation (1:5 reduction). A novel method for representation and recognition of 3D Objects/Targets based on 3D frequency domain representation was also developed and tested.

14. SUBJECT TERMS	15. NUMBER OF PAGES
AUTOMATIC TARGET RECOGNITION, FACE RECOGNITION, 3D OBJECT REPRESENTATION, SPECTRAL SIGNATURES	43
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unclassified

ONR Annual Productivity Report for 30 Sep. 1996 to 31 July. 1997
DARPA/ONR Grant No. N00014-93-1-1088

Submitted to:

DR. THOMAS M. MCKENNA
OFFICE OF NAVAL RESEARCH
ONR 342 BALLSTON TOWER ONE
800 NORTH QUINCY STREET
ARLINGTON, VA 22217-5660

email: mckennt@onrhq.onr.navy.mil
FAX: (703) 696-1212 or (703) 696-0103
Voice: (703) 696-4503

1. **Principal Investigator Name:** Jezekiel Ben-Arie
2. **Institution:** University of Illinois at Chicago
3. **Full Address:** EECS Dept. (M/C 154), 851 S. Morgan St., Chicago IL 60607.
4. **Current Phone:** (312) 996-2648
5. **Current Fax:** (312) 996-2648
6. **email:** benarie@eecs.uic.edu
7. **Project Title:** Automatic Target Recognition by Non-Orthogonal Image Expansion with Neurocomputers
8. **Do you post web pages/ftp site that provides reports and descriptions of research?**
   FTP Site: Yes URL: **ftp://hezi.eecs.uic.edu/pub/papers/**
9. **For the current fiscal year, please provide:**
   a. **Number of ONR supported:**
      i. **Papers published in refereed journals:** 1
      ii. **Papers accepted for publication in refereed journals:** 2
      iii. **Papers submitted to refereed journals:** 4
      iv. **Papers or reports in non-refereed journals:** 0
      v. **Books or book chapters published:** 0
      vi. **Books or book chapters in press:** 1
      vii. **Papers in refereed conferences:** 6

*** Attach list of papers and other publications with full citation, arranged by class i through vii.***
Number of Presentations: Invited: 1 **Contributed:** 6

   b. **Number of ONR supported patents/inventions filed:** None **or granted:** None.
      *** Attach title and brief description of patents/inventions, if any.***

c.

	Trainee Data	Total	Female	Male	Minority	Non-US Citizens
i	No. of Grad. Students	5	0	5	0	4
ii	No. of PostDoctorals	0	0	0	0	0
iii	No. of Undergraduates	1	0	1	0	1

d. **Number, cost and description of equipment items costing more than $1,000 that were purchased on your ONR grant:** None.

e. **Awards/Honors to PI and/or to members of PI's research group (please describe):**
Dibyendu Nandy (Ph.D. student) received the **Andrew S. Foundation Fellowship** for the year 1996. Dr. Raghunath K. Rao, (Post-Doctoral Research Associate funded by the ONR grant), was elected to the **Who's Who Amongst Students in American Universities and Colleges for Outstanding Merit and Accomplishment.** In addition, Dr. Rao, along with Dibyendu Nandy and Zhiqian Wang (Ph.D. student) were elected as **Associate Members of Sigma Xi, The Scientific Research Society.**

f. **Brief description of all transitions (or intended transitions) of your ideas or techniques to industry, to military laboratories or to military application. Include the name of the organization to which the work was transitioned, and the name of the point of contact. (Note: Transitions are of increasing importance in the current applications-oriented climate of DoD research):**

1. The Expansion Matching (EXM) method which is a general method for matching signal/image templates and is proven superior over correlation matching, has been widely published in major journals, books, and international and domestic conferences. Many requests for reprints of this literature have been received, and papers have been widely distributed. A public-domain implementation of the multiple-template Expansion Matching has been developed and delivered to researchers in universities, military and commercial organizations, such as Army Missile Command, NASA, Boston Univ., Univ. of South Florida, Hewlett Packard, Datacube, SYS-TECH Solutions, Inc., etc. An anonymous FTP site for dissemination of EXM software and literature has been setup.

   Expansion Matching (EXM) provides a new and powerful tool that can be used in a wide variety of applications, such as recognition, stereo, motion estimation, image flow. In fact any task that uses correlation matching is significantly improved by using Expansion Matching. EXM yields improved matching discrimination (usually about 25dB better than normalized correlation) and better localization with substantially decreased spurious responses.

   The address for the anonymous FTP site is hezi.eecs.uic.edu, which contains the EXM software implementation in the file:

/pub/arpa/exm/exm.tar.gz. Also, the software implementation for the EXM-based optimal edge detection system is available in the file: /pub/arpa/edge/edge.tar.gz. Literature on DARPA-funded research (including EXM, the shape description network, the recent affine invariant object recognition network and the generalized feature extraction scheme) is available in the directory: /pub/papers.

2. A new method for iconic object/target recognition based on novel Affine Invariant Spectral Signatures (AISS) was recently developed. This method achieves for superior results over other methods. Recognition rates of 100% were achieved in a wide range of viewing directions, with a large library of models. The method is robust not only to viewing directions and scale but also to noise/clutter and reduced resolutions.

The point of contact for any questions is :

Jezekiel Ben-Arie
Associate Professor
University of Illinois at Chicago
EECS Dept. (M/C 154)
851 S. Morgan St.
Chicago, IL 60607
Phone: (312) 996-2648
Fax: (312) 996-2648
Email: benarie@eecs.uic.edu

List of ONR Supported Publications for 1996-97

i. Papers in Refereed Journals: 1

1. Wang, Z., and Ben-Arie, J., "Optimal Ramp Edge Detection using Expansion Matching," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 43, No. 11, November 1996, pp. 1093-1105.

ii. Papers Accepted for Publication in Refereed Journals: 2

1. Ben-Arie, J. and Wang, Z., "Iconic Representation and Recognition Using Affine-Invariant Spectral Signatures," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, accepted May 1997.
2. Nandy, D., Wang Z. and Ben-Arie J., "Generalized Feature Extraction using Expansion Matching and Karhunen-Loeve Basis Functions," **IEEE Transactions on Image Processing**, accepted pending revision June 1997.

iii. Papers Submitted for Publication in Refereed Journals: 4

1. Ben-Arie, J. and Nandy, D., "Volumetric/Iconic Frequency Domain Representation for Objects with application for Pose Invariant Face Recognition," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, submitted January 1997.
2. Wang, Z. and Ben-Arie, J., "Hierarchical Shape Description and Similarity-Invariant Recognition Using Gradient Propagation," **International Journal of Computer Vision**, submitted March 1997.
3. Wang, Z. and Ben-Arie, J., "Shape Description and Invariant Recognition Employing Connectionist Approach," **Signal Processing**, submitted March 1997.
4. Shroff, H.K., and Ben-Arie, J., "Parallel Methods for Axial Shape Description based on Magnetic Field and Gradient Propagation," **IEEE Transactions on Image Processing**, submitted March 1997.

iv. Papers Accepted for Publication in Non-Refereed Journals: None

v. Books or Book Chapters Published: None

vi. Books or Book Chapters in Press: 1

1. Ben-Arie, J., and Wang, Z., "Gabor Kernels for Affine-Invariant Target Recognition," To appear in **Gabor Analysis and Algorithms: Theory and Applications**, (Birkhauser Press) H. G. Feichtinger and T. Strohmer (Eds.), 1997.

vii. Papers in Refereed Conferences: 6

1. Ben-Arie, J. and Wang, Z., "SVD and Log-Log Frequency Sampling with Gabor Kernels for Invariant Pictorial Recognition," **IEEE Signal Processing Society 1997 International Conference on Image Processing, (ICIP'97)**, Santa Barbara, CA, October 26-29, 1997.

2. Ben-Arie, J. and Nandy, D., "Using the Fourier Slice Theorem for Representation of Object Views and Models with application to Face Recognition," **IEEE Signal Processing Society 1997 International Conference on Image Processing, (ICIP'97)**, Santa Barbara, CA, October 26-29, 1997.

3. Ben-Arie, J. and Wang, Z., "Pictorial Recognition Using Affine-Invariant Spectral Signatures," **IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)**, San Juan, Puerto Rico, June, 1997, pp. 34-39.

4. Ben-Arie, J. and Nandy, D., "Representation of Objects in a Volumetric Frequency Domain with application to Face Recognition," **IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)**, San Juan, Puerto Rico, June, 1997, pp. 615-620.

5. Ben-Arie, J., Wang, Z. and Rao, K. R., "Iconic recognition with affine-invariant spectral signatures," in **Proceedings of the IAPR-IEEE 13th International Conference on Pattern Recognition (ICPR'96)**, Vol. 1, pp. 672-676, Vienna, Austria, Sept. 1996.

6. Wang, Z., Rao R. K., Nandy, D. Ben-Arie J., and Jojic N., "A Generalized Expansion Matching Based Feature Extractor," in **Proceedings of the IAPR-IEEE 13th International Conference on Pattern Recognition (ICPR'96)**, Vol. 2, pp. 29-33, Vienna, Austria, Sept. 1996.

Final Report
DARPA/ONR Grant No. N00014-93-1-1088

PI: Jezekiel Ben-Arie
Associate Professor, EECS Dept. (M/C 154)
University of Illinois at Chicago
851 S. Morgan St., Chicago, IL 60607
URL/FTP: ftp://hezi.eecs.uic.edu/pub/papers/

September 20, 1997

Abstract

This report describes two novel approaches to pose invariant object representation and recognition. The first section describes an efficient approach to pose invariant pictorial object recognition employing spectral signatures of image patches that correspond to object surfaces which are roughly planar. A complete affine invariance of the signatures is achieved by a log-log sampling configuration in the frequency domain. Based on Singular Value Decomposition (SVD), the affine transform is decomposed into slant, tilt, swing, scale and 2D translation. Unlike previous log-polar representations which were not invariant to slant (i.e. foreshortening only in one direction), our new configuration yields complete affine invariance. The proposed log-log configuration can be employed both globally or locally by Fourier or Gabor transforms. A novel model based affine invariant segmentation scheme enables to isolate and recognize several objects in cluttered images. The actual signature recognition and 3D pose estimation is performed by multi-dimensional indexing in a pictorial dataset represented in the frequency domain. Experimental results with a dataset of 26 models show 100% recognition rates in a wide range of 3D pose parameters and imaging degradations: $0-360°$ swing and tilt, $0-82°$ of slant (more than 1:7 foreshortening), more than 3 octaves in scale change, window-limited translation, high noise levels (0 dB) and significantly reduced resolution (1:5).

In the second section, a novel method for representing 3-D objects that unifies viewer and model centered object representations is presented. A unified 3-D frequency-domain representation (called Volumetric/Iconic Spectral Signatures - V/ISS) encapsulates both the spatial structure of the object and a continuum of its views in the same data structure. The frequency-domain image of an object viewed from any direction can be directly extracted employing an extension of the Projection Slice Theorem, where each Fourier-transformed view is a planar slice of the volumetric frequency representation. The V/ISS representation is employed for pose-invariant recognition of complex objects such as faces. The recognition and pose estimation is based on an efficient matching algorithm in a four dimensional Fourier space. Experimental examples of pose estimation and recognition of faces are also presented.

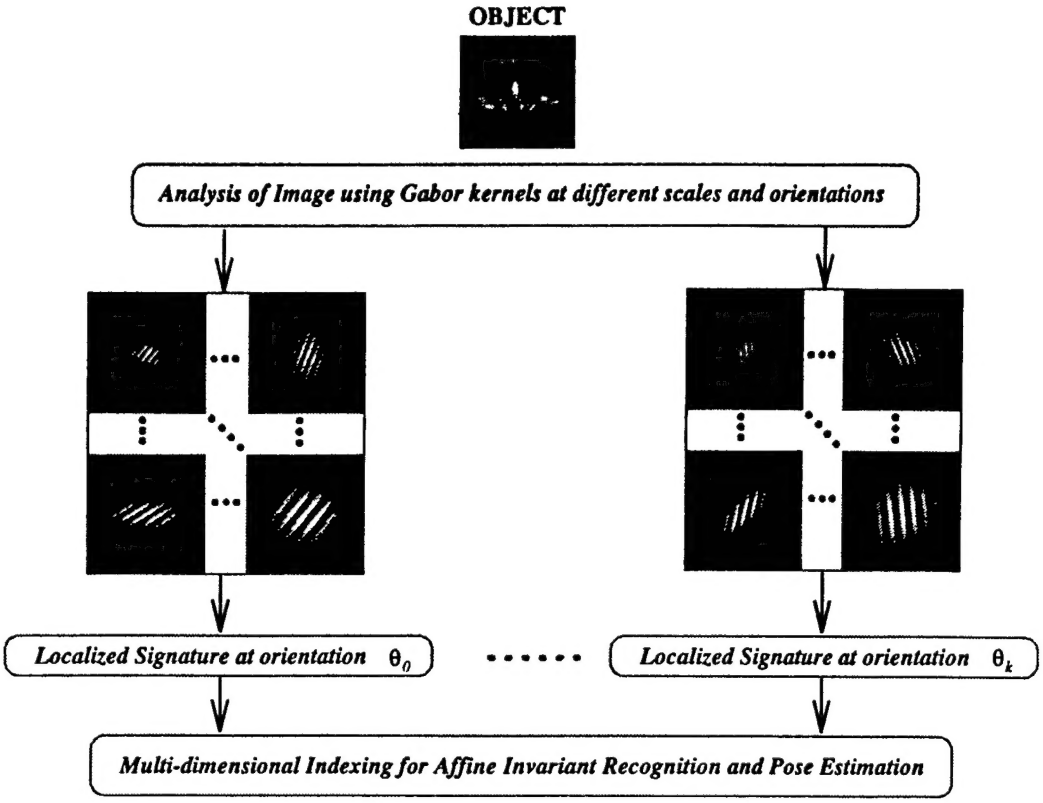


Figure 1: Block diagram of pictorial recognition scheme.

1 Pictorial Recognition of Objects Employing Affine Invariance in the Frequency Domain

1.1 Introduction

This section describes a method for pose invariant pictorial recognition of 3D objects employing frequency domain techniques. By the term pictorial recognition we mean that the recognition is achieved by matching in feature space a given image to a model dataset which consists of various object pictures. Even though such a pictorial model dataset contains only few aspects for each 3D object represented, it is still possible to achieve robust recognition of objects in a wide range of viewing directions and distances - if one employs pose invariant matching methods as illustrated in this section. Hence, the number of pictures required to represent an object in the database could become quite small.

If we treat pixel values as real numbers, we can regard each picture of an object instance as a point in R^M, where M is the number of pixels in the picture. As the parameters of the object's pose vary, the point in the M-dimensional space R^M traces out a l-dimensional manifold, where

l is the number of pose parameters. Additional parameters that relate to illumination and sensor characteristics may further increase the l-dimensionality. Different objects generate different manifolds in R^M. In this setting, object recognition can be posed as finding the closest point of any of the manifolds to a given test image. If the point is close enough to one of the manifolds, we can claim that the given test images belongs to a particular object whose manifold was the one matched.

The problem with this approach is that such setting requires construction of a large number of very complex manifolds which correspond to changing views of objects as a function of their pose in space. A drastic simplification is necessary to render such an approach practically implementable. One feasible suggestion is to find a pictorial representation which is invariant to the largest number of pose parameters. For each parameter eliminated, we can reduce the dimensionality and simplify the overall representation.

Many approaches have been suggested in the area of invariant pictorial representation and recognition. Best fitting to the real problem are methods employing perspective projection invariants. Such is the work of Jacobson and Wechsler [41] who employed 4D Wigner distribution [41] combined with back projection to achieve perspective invariance in 6 dimensional search space. Since perspective invariance leads to unmanageable complexity (of 4D correlation in 6D search space), it is advantageous to approximate the perspective transformation by simple ones such as the affine transformation. Although affine transformation is only an approximation of perspective transformation, it reflects quite accurately the real 3D geometric distortions of a planar object when the dimensions of the object are relatively small compared with the distance between the imaging system and the object itself. Several previous works suggested affine invariant recognition of planar objects based on invariant moments [20] [19] and contours [39] [8]. In real imagery, both types of methods require accurate segmentation and edge grouping and therefore they are quite sensitive to illumination, noise, clutter, partial occlusion and perspective geometrical distortions. On the other hand, our approach which is based on representing the pictorial dataset in the frequency domain has few advantages. First, it allows to eliminate planar translation effects in the imaging plane by considering only the magnitude of the Fourier (or Gabor) transform. Second, the representation of noise and clutter in the frequency domain can be easily filtered and removed. And third, as demonstrated in Section 1.4.2, the frequency based signatures are quite tolerant to distortions that arise from inaccurate segmentation, multiplicative illumination effects and the actual perspective imaging.

Once the planar translation effects are removed from the representation, the next task is to achieve invariance to the other pose parameters, i.e. the three rotational degrees of freedom and the remaining translation parameter, i.e. translation normal to the imaging plane (translation

along the optical axis). In Section 1.2 and in [7] [6] [5] [1] [3], we show that if we limit ourselves to pose-invariant recognition of planar objects and surfaces, the above parameters can be represented by slant, tilt, swing and scale parameters[1]. By the term slant, we refer to the angle between the normals to the image plane and the object-plane. The tilt is defined as the angle between the X-axis in the image plane and the axis of intersection of the object-plane with the image plane (tilt axis). In an orthographic projection of planar shapes, slant causes foreshortening only along the normal to the tilt axis in the image plane, while distances along the tilt axis remain unaltered. Previous approaches developed for pictorial recognition which are based on log-polar representations in the frequency domain [14] [11] [43] [22] or in the spatial domain [40] are not invariant to uneven distortion caused by foreshortening. The log-polar configuration is invariant only to scale and rotation, i.e. similarity transformation. However, the similarity transform is only a subset of the complete affine transform and cannot represent all the geometric distortions caused by orthographic projection.

In this section, an affine-invariant representation is achieved by sampling the frequency domain representation in a novel configuration which is logarithmic in two orthogonal axes, i.e log-log configuration. As elaborated Section 1.2 the log-log configuration is invariant to translation, slant and scale. Invariance to the remaining degrees of freedom i.e. to tilt and swing (rotation around the optical axis) is attained by a union of swung log-log configurations. As described in Section 1.2 and in [7] [6] [5], it is feasible to derive the spectral signatures by several methods that include short-term Fourier transform, Gabor transform and also two dimensional Gaussian derivatives. All these methods are intended to obtain a spatially local representation of image patches in the frequency domain. Local representations enable to independently recognize several image patches in the same image. Hence, an object which is composed of several roughly planar surfaces can be robustly recognized by recognizing a few of its surfaces or parts.

We choose to use the Gabor kernels since Gabor functions yield the smallest conjoint space-bandwidth product permitted by the uncertainty principle of Fourier analysis [17] [41]. This allows us to derive local frequency characteristics of image patches since Gabor kernels form a complete basis for signal representation. Since the local Gabor signature obtained is still sensitive to location of the centers, we develop in Section 1.4.2 a model based affine invariant segmentation method. This approach enables to segment image regions with predetermined shape (rectangular, circular etc.) with any affine distortion. The segmentation method is based on image convolution with a set of basis functions derived by Karhunen-Loeve (K-L) transform. To achieve more accurate segmentation, an additional stage of flexible matching is also included. The signature

[1]In this section, we use the terms "slant" and "tilt" to denote plane rotations in orthographic projection. To avoid confusion, we comment that these terms are usually employed in the context of perspective projections.

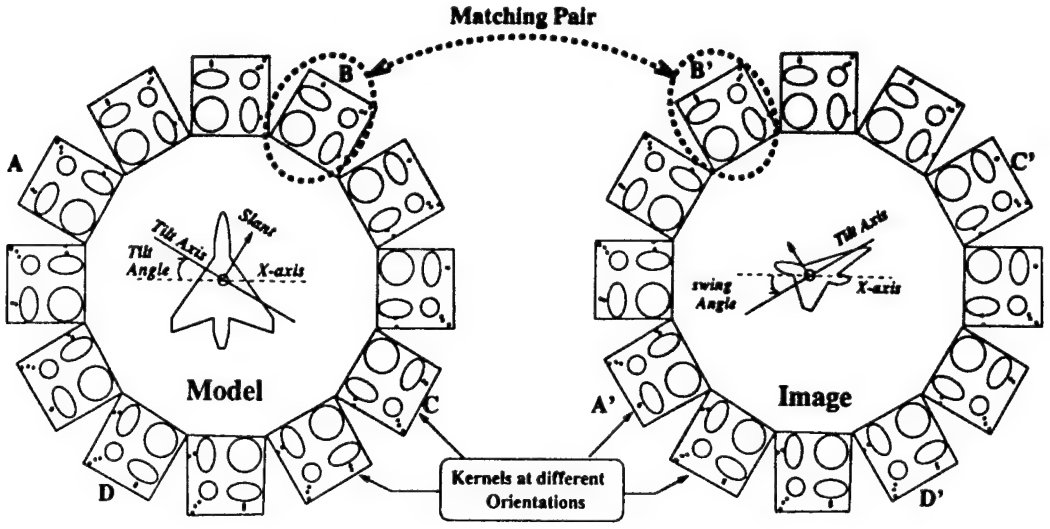


Figure 2: Invariance of signatures to swing and tilt of the shape. For the displayed swing and tilt of the airplane shape, the oriented kernel pairs **A-A'**, **B-B'**, **C-C'**, and **D-D'** yield invariant signatures. Note that since the kernels have symmetry properties, it is required to implement only one quadrant of kernels spanning 90 degrees of orientation to achieve 360 degrees of swing and tilt invariance.

of each segmented region is then derived independently and objects can be recognized even in cluttered scenes as demonstrated in Section 1.4.2.

In Section 1.2 we provide a mathematical description of the affine invariant representation both in the spatial and frequency domains. In Section 1.3 we describe the recognition techniques and in Section 1.4 we illustrate the experiments which achieve quite a robust recognition in a wide range of viewing conditions.

1.2 Affine-Invariant Spectral Signatures (AISSs)

Our overall approach is based on pictorial recognition of image patches that correspond to object surfaces that are approximately planar. As elaborated later, object surfaces can be recognized in a general 3D pose. The class of objects that can be recognized is not limited to convex objects and also includes concave objects or objects with holes, etc. As long as an object has at least one approximately planar surface with distinctive features, it may be recognized by this approach. As experimental results demonstrate in Section 1.4, many non-planar objects which have approximately flat shapes such as hands, airplanes, etc. are robustly recognized with our approach as well.

We use the affine transformation to simulate transformation of a planar shape that undergoes 3D rotation and 3D translation, and is then orthographically projected onto the image plane and

scaled (reduced or increased in size). A point $\underline{X} = (x, y)^T$ in the coordinate system of the shape is affine-transformed to a point in imaging plane's coordinate system $\underline{X}_a = (x_a, y_a)^T$ according to the following formula:

$$\begin{pmatrix} x_a \\ y_a \end{pmatrix} = C \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} g \\ h \end{pmatrix} \tag{1}$$

where matrix C represents tilt and slant operation, $(g, h)^T$ denotes translation. This general formulation represents any orthographic projection plus scaling of planar shapes. Such a projection approximates perspective projections quite accurately if the viewing distance of the shape is relatively large with respect to the shape's dimensions. Based on Singular Value Decomposition (SVD), the matrix C can be decomposed as follows:

$$\begin{pmatrix} c_{11}\, c_{12} \\ c_{21}\, c_{22} \end{pmatrix} = \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} \cos(\tau) & -\sin(\tau) \\ \sin(\tau) & \cos(\tau) \end{pmatrix} \tag{2}$$

where λ_1 and λ_2 are eigenvalues of CC^T, ϕ and τ are angles related to eigenvectors of CC^T and C^TC. In practical situations, C is usually a nonsingular matrix, so λ_1 and λ_2 have positive values. If we arrange the eigenvalues so that $\lambda_1 \geq \lambda_2$, the eigenmatrix Λ can be posed as

$$\Lambda = \sqrt{\lambda_1} \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{\lambda_2/\lambda_1} \end{pmatrix} \tag{3}$$

According to Eq. (1) and Eq. (2), any orthographic projection of points on a plane can be represented by a sequence of transformations which include translation, tilt[1], slant, scale and swing (rotation around the optical axis). To represent 3D rotation of a plane it is necessary to use slant and tilt transformations in which the shape is posed on a plane which is slanted and tilted with respect to the imaging plane. Slant angle is measured between the normals of the imaging and shape planes. Tilt is defined as the angle between the the X-axis in the imaging plane and line L created by the intersection of the imaging and shape planes (the tilt axis L). Here, this angle is defined as τ in Eq. (2). ϕ in Eq. (2) represents shape swing within the imaging plane. The slant angle σ corresponds to shape foreshortening in the imaging plane along the axis normal to the line L. The foreshortening ratio is equal to $\cos(\sigma) = \sqrt{\lambda_2/\lambda_1}$. In contrast to slanting, scaling causes uniform foreshortening (or enlargement) in the imaging plane in all directions. The scale factor is equal to λ_1 in Eq. (3). The above parameters of translation, swing, scale, slant and tilt completely represent the scaled orthographic projection.

When a planar object undergoes affine transformation, the frequency spectrum of its image is also transformed by a similar set of transformations. Given a function $f(\underline{X})$ with Fourier

Transform as $\mathcal{F}(u,v)$, its affine transformed version $f_a(X) = f(C^{-1}X - C^{-1}(g,h)^T)$ has the frequency spectrum as follows:

$$|\mathcal{F}_a(u_a,v_a)| = |C||\mathcal{F}(u,v)|$$
$$\begin{pmatrix} u_a \\ v_a \end{pmatrix} = \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix} \begin{pmatrix} {\lambda_1}^{-1/2} & 0 \\ 0 & {\lambda_2}^{-1/2} \end{pmatrix} \begin{pmatrix} \cos(\tau) & -\sin(\tau) \\ \sin(\tau) & \cos(\tau) \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \tag{4}$$

Thus, the effect of affine transformation on the spectrum is almost the same as that of the affine transform on the object in spatial domain except for two major differences: First, the spectrum is inversely scaled and slanted. Secondly, shape translations parallel to the image plane do not affect the spectrum.

The coordinate transformation in Eq. (4) can be rewritten as

$$\begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix} \begin{pmatrix} u_a \\ v_a \end{pmatrix} = \begin{pmatrix} {\lambda_1}^{-1/2} & 0 \\ 0 & {\lambda_2}^{-1/2} \end{pmatrix} \begin{pmatrix} \cos(\tau) & -\sin(\tau) \\ \sin(\tau) & \cos(\tau) \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \tag{5}$$

From Eq. (5), sampling the affine transformed shape's spectrum $|\mathcal{F}_a(u_a,v_a)|$ along two orthogonal directions at angles ϕ and $\phi+\pi/2$ results in a spectral representation we call spectral signature $N_a(\omega_1,\omega_2,\phi)$, where $\omega_1 = \log_r(|u_a\cos(\phi) - v_a\sin(\phi)|)$ and $\omega_2 = \log_r(|u_a\sin(\phi) + v_a\cos(\phi)|)$. Sampling the original shape's spectrum $|\mathcal{F}(u,v)|$ along two orthogonal directions at angles τ and $\tau+\pi/2$ results in the model's spectral signature $N(\omega_1,\omega_2,\tau)$, where $\omega_1 = \log_r(|u\cos(\tau) - v\sin(\tau)|)$ and $\omega_2 = \log_r(|u\sin(\tau) + v\cos(\tau)|)$. The two spectral signatures thus derived are related as $N_a(\omega_1,\omega_2,\phi) = |C|N(\omega_1-\alpha_1,\omega_2-\alpha_2,\tau)$, where $\alpha_1 = \log_r(\sqrt{\lambda_1})$ and $\alpha_2 = \log_r(\sqrt{\lambda_2})$. We note that the signature is not altered due to slanting and scaling but only is translated in (ω_1,ω_2) plane (see in Fig. 4 and Fig. 5).

It is noted here that a 2D Cartesian version of the Mellin transform - implemented here in the frequency domain - which is defined as

$$M_a(\xi_1,\xi_2,\phi) = \iint |\mathcal{F}_a(u_a\cos(\phi) - v_a\sin(\phi), u_a\sin(\phi) + v_a\cos(\phi))|$$
$$(u_a\cos(\phi) - v_a\sin(\phi))^{-j\xi_1-1}(u_a\sin(\phi) + v_a\cos(\phi))^{-j\xi_2-1}\,du_a\,dv_a \tag{6}$$

also achieves invariance to slanting and scaling which result in linear phase shifts proportional to $\ln(\sqrt{\lambda_1})$ and $\ln(\sqrt{\lambda_2})$.

$$M_a(\xi_1,\xi_2,\phi) = |C|\left(\frac{1}{\sqrt{\lambda_1}}\right)^{-j\xi_1}\left(\frac{1}{\sqrt{\lambda_2}}\right)^{-j\xi_2} M(\xi_1,\xi_2,\tau) \tag{7}$$

$M_a(\xi_1,\xi_2,\phi)$ is the Mellin transform of $|\mathcal{F}_a(u_a,v_a)|$ with axis direction at ϕ. $M(\xi_1,\xi_2,\tau)$ is the Mellin transform of the original spectrum $|\mathcal{F}(u,v)|$ with axis direction at τ.

The estimation of slant and scale parameters from the relative phase shift between $M_a(\xi_1, \xi_2, \phi)$ and $M(\xi_1, \xi_2, \tau)$ is quite difficult. On the other hand, these parameters are easier to estimate from the relative shift between our spectral signatures $N(\omega_1, \omega_2, \tau)$ and $N_a(\omega_1, \omega_2, \phi)$ in (ω_1, ω_2) plane. Hence, the signature $N_a(\omega_1, \omega_2, \phi)$ from the affine transformed object is a shifted version of the signature $N(\omega_1, \omega_2, \tau)$ from the object itself except for a scalar $|C|$. The shift in the 2D signature plane (ω_1, ω_2) directly depends on the slant and the scale included in the affine transformation. In order to account for the remaining two rotational degrees of freedom, i.e. swing and tilt, we generate for the affine transformed object a set of signatures $\{N_a(\omega_1, \omega_2, \theta_1); 0° < \theta_1 < 360°\}$ which have equally spaced orientations and which span the range of 360 degrees. A set of signatures $\{N(\omega_1, \omega_2, \theta_2); 0° < \theta_2 < 360°\}$ for the model are also created in the same way. Among the set of pictorial signatures generated, there exists one which matches the signature of the model object except for a translation in the (ω_1, ω_2) plane - which represents scale and slant differences.

Figure 1 displays a block diagram of the overall system. The image is correlated with a set of Gabor kernels. The frequencies of the kernels are derived from a logarithmic sampling according to Eq. (8) and Eq. (9). This set is centered at various 'interest locations' which correspond to approximate centers of prominent image patches[2]. A set of spectral signatures is then generated. Each signature represents a local image patch. These signatures are then independently recognized using Multidimensional Indexing (MDI). The 3D pose (slant, tilt, swing and scale) of each recognized patch is also obtained as a by-product.

The affine-invariant representation presented in this section is based on a set of elliptical 2D Gabor kernels defined as

$$g_{mn}^{f_x, f_y, \theta_l}(x, y) = e^{-j2\pi\left(\frac{x_l f_x}{\sigma_{X_m}} + \frac{y_l f_y}{\sigma_{Y_n}}\right)} \cdot e^{\left(-\frac{x_l^2}{2\sigma_{X_m}^2} - \frac{y_l^2}{2\sigma_{Y_n}^2}\right)}$$
$$\begin{pmatrix} x_l \\ y_l \end{pmatrix} = \begin{pmatrix} \cos\theta_l & \sin\theta_l \\ -\sin\theta_l & \cos\theta_l \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{8}$$

where f_x, f_y are frequency coefficients, $f_x, f_y = 1...N_f$. The standard deviations σ_{X_m} and σ_{Y_n} of these elliptical kernels vary in a geometrical progression with the indices m and n as

$$\sigma_{X_m} = \gamma^{m-1}\sigma_0;\ \sigma_{Y_n} = \gamma^{n-1}\sigma_0;\ m, n = 1...N_\sigma \tag{9}$$

where the geometric ratio $\gamma > 1$ and the smallest standard deviation σ_0 are constants. The indices m and n define a signature space (m, n) and also determine the sampling points in the (ω_1, ω_2) plane for a given set of σ_0, f_x and f_y . In addition, the Gabor in Eq. (8) is modulated in two orthogonal axes (which have orientation θ_l denoted by X_l and Y_l) by a complex sinusoid

[2]As described in Section 1.4.2, these patches are first segmented from the image and the signature obtained is thus not sensitive to the exact locations of the center nor to neighboring image regions

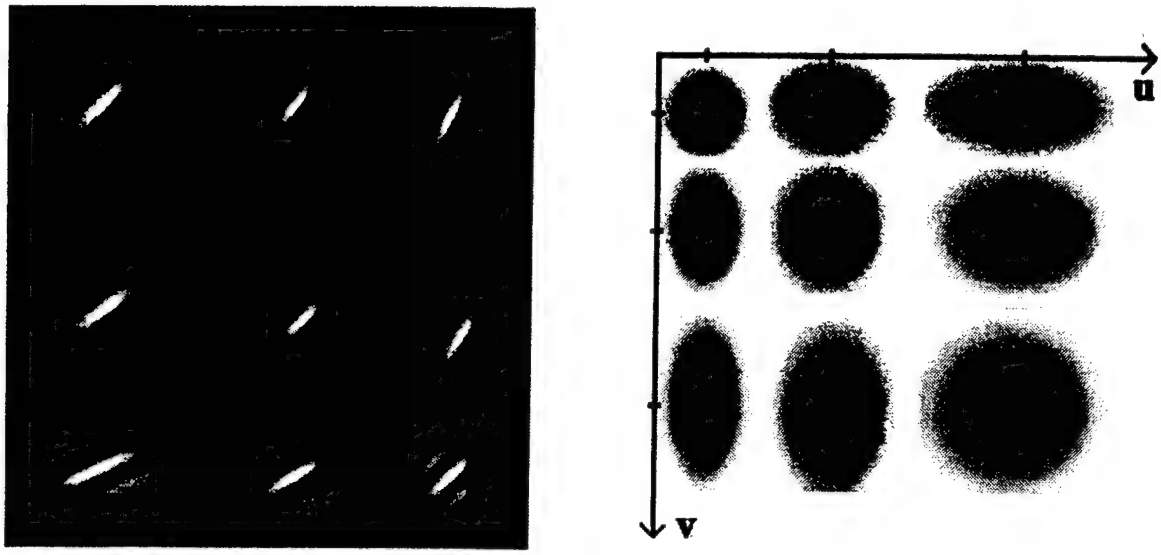


Figure 3: Partial subset of kernels K^{θ_l} ($f_x = 1$ and $f_y = 1$) for orientation $\theta_l = 0$ degrees in spatial domain (left) and the configuration of corresponding kernels in frequency domain (right). Each subset K^{θ_l} completely spans the frequency domain.

with periods proportional to the deviations of corresponding Gaussian profile. The parameter θ_l denotes the orientation of the kernel and uniformly spans the range $[0, 360)$ degrees in discrete steps $l = 1...N_\theta$.

The above scheme generates a subset of modulated Gaussian kernels $K^{f_x,f_y,\theta_l} = \{g_{mn}^{f_x,f_y,\theta_l}$; $m, n = 1...N_\sigma\}$ with identical orientation θ_l and identical frequency coefficients (denoted by f_x and f_y), but with varying aspect ratio and size (indexed by σ_{X_m} and σ_{Y_n}). For each orientation θ_l, we have a cumulative subset K^{θ_l} of kernels which includes all the frequency coefficients, i.e., $f_x, f_y = 1...N_f$. The complete set of kernels K consists of the union of all the subsets K^{θ_l} swung to different orientations θ_l ; $l = 1..N_\theta$ that uniformly span 360 degrees. In practice it is required only to generate kernels that span one quadrant (90 degrees) of orientation. All the other kernels can be constructed from this reduced set using symmetry properties. An example of one subset of kernels K^{θ_l} with $\theta_l = 0$ degrees is illustrated in Fig. 3 (left, only the real parts of the kernels). The frequency spectrum of this subset of kernels is also illustrated in Fig. 3 (right), and shows that each subset of kernels K^{θ_l} completely spans the band-limited frequency domain of interest and is logarithmically spaced as needed.

When a local image patch $I(x, y)$ is correlated with this configuration of kernels, it generates a set of multi-dimensional spectral signatures $\{S^{f_x,f_y,\theta_l}; f_x, f_y = 1...N_f$, $l = 1...N_\theta\}$ composed of the correlation (projection) coefficients of all the kernels. Explicitly,

$$S^{f_x,f_y,\theta_l}(\sigma_{X_m}, \sigma_{Y_n}) = | < g_{mn}^{f_x,f_y,\theta_l}(x, y), I(x, y) > | \qquad (10)$$

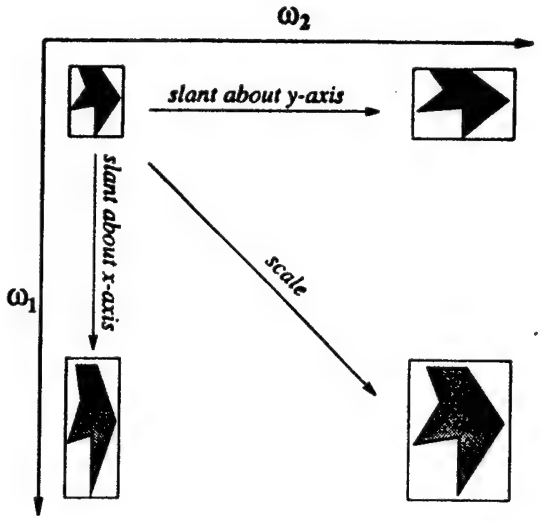


Figure 4: Shifting property of the spectral signature in the (ω_1, ω_2) plane with respect to scaling and slanting of an arbitrary shape.

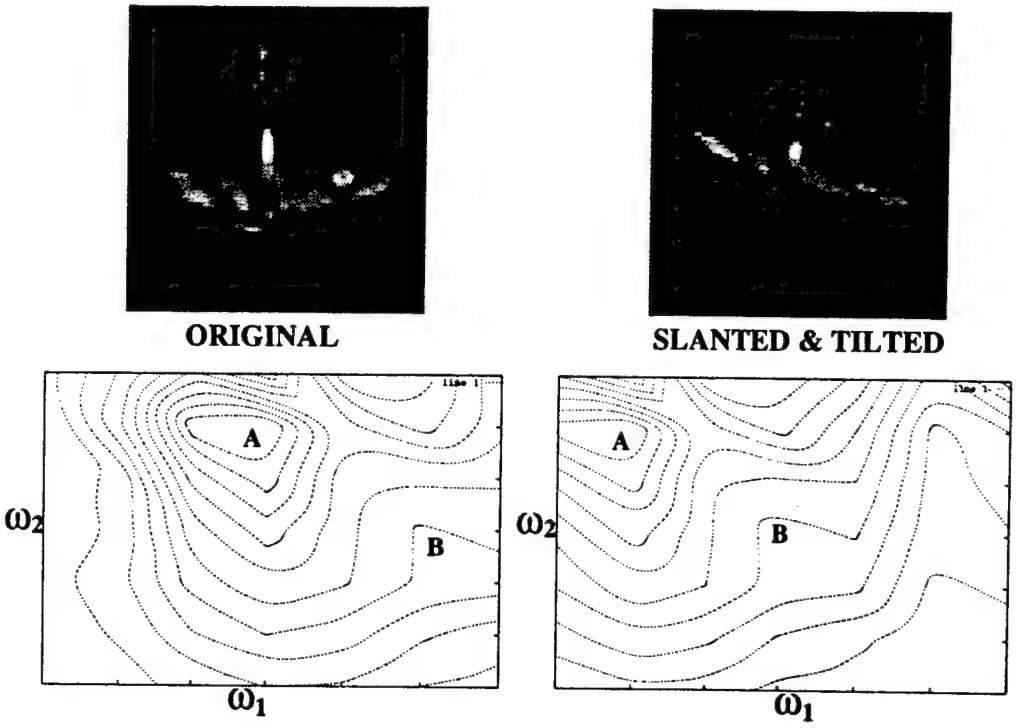


Figure 5: Contour plots of the signature for the original object (left) and shifted spectral signature obtained for the slanted and tilted object (right). The contour plots demonstrate invariance to slant and tilt except for a shift in the signature space. The labels **A** and **B** illustrate the shift in the signature.

for $m, n = 1...N_\sigma$, where $|\cdot|$ denotes modulus of a complex number.

The property of slant-invariance arises from the fact that when the image patch corresponds to a slanted shape, say in axis X_l, its signature S^{f_x,f_y,θ_l} shifts in the direction of the slant, i.e. ω_2, with respect to the signature of the unslanted shape. Also, when the shape is scaled, all the signatures $\{S^{f_x,f_y,\theta_l}\ ;\ l = 1...N_\theta\}$ shift equally, i.e. diagonally, in the (ω_1, ω_2) plane. Hence, any combined slant and scale results in a corresponding shift in the (ω_1, ω_2) plane. Fig. 4 illustrates these properties of the signature. Fig. 5 displays contour plots of the signature of the airplane model and the corresponding signature when the airplane is slanted by 60 degrees with a tilt of 15 degrees. It is easily observed (see the labels **A** and **B** on the plots displayed for easy registration) that the signature does not change except for a translation in the (ω_1, ω_2) plane. The translation between a model signature and the image signature can be used to compute the relative 3D pose between the two. The difference in ω_1 and ω_2 can be directly translated into relative slant and scale. The other angular pose parameters of tilt and swing can also be retrieved as described below. The X and Y coordinates are derived from the image, and the depth parameter can be derived from the scale.

Since shapes can be slanted and tilted in any orientation in space, one has to generate a subset of kernels for each tilt direction and for each orientation, which forms two rotational degrees of freedom. These two degrees of freedom are dealt with by using the complete set of kernels K both for the model signature and for the image signature. This is demonstrated in Fig. 2, where it is shown that even if the model is tilted and swung, there is exact correspondence between four of the model signatures (marked by labels **A** through **D**) and four of the image signatures (marked by labels **A'** through **D'**). This invariance to swing and tilt is possible only because both the model and the image are processed by subsets of kernels at different orientations. In Section 1.4, it is experimentally found that sampling of 7.5 degrees in θ_l achieves a sufficient interpolation to accommodate any intermediate values of tilt and swing.

From Eq. (10), we see that the signatures are related only to the magnitudes of the complex correlation coefficients, the phase information being completely eliminated. Thus, the signatures obtained are - to a large extent - invariant to limited translation of the object within the localized Gabor support (approximately $\pm\sigma$).

Hence, the combined set of kernels K, composed of all the subsets K^{θ_l} sufficiently covers scale, slant, tilt, swing, and translation, i.e. all affine transformation parameters that simulate the scaled orthographic projection.

1.3 Affine-Invariant Recognition by Multi-Dimensional Indexing

Our recognition scheme is based on the affine-invariant nature of the spectral signatures described in Section 1.2. As explained above, when the shape is slanted with a tilt axis of orientation θ_l, the signature S^{f_x,f_y,θ_l} - which corresponds to the tilt orientation θ_l - undergoes simple shifts in the (ω_1, ω_2) plane that correspond to scale and slant transformations. The purpose of the indexing scheme is to robustly identify the image patch from its set of signatures. Each signature S^{f_x,f_y,θ_l} corresponds to a combination of orientation θ_l and the frequency coefficients f_x, f_y. A robust recognition scheme is required since the signatures could be partially distorted due to illumination variations, due to the discrete nature of the orientation or due to the limited range of scales. Furthermore, irrelevant clutter in the receptive field and partial occlusion can result in additional distortions.

In order to overcome these signature distortions, we implement a voting scheme using the spectral signatures, based on MDI [13]. MDI basically relies on the same principles as the geometric hashing method [28]. The main difference is that the indices for the hash table have few dimensions which correspond to few invariant shape characteristics. The low dimensionality of geometric hashing causes overcrowding of bins, and the hash table sometimes saturates even with a small number of objects. On the other hand, MDI improves the robustness of the recognition (which is expressed as the ratio of the highest vote to the next highest vote). This result was also observed by [34]. The innovation of our indexing scheme is that it is implemented in the frequency domain using spectral signatures. Additional merits of MDI are that the retrieval size of the database is considerably increased, the overcrowding of bins in the hash table is almost eliminated, and coarser quantization can be used without reducing discrimination. We experimentally found that the large dimensionality in the indexing space does not significantly increase the search times.

In our indexing scheme, the hash table is updated by each model using all its signatures S^{f_x,f_y,θ_l}. 11-dimensional indices are generated for the models (to each index, an additional nine dimensional information vector is also attached). Every index corresponds to a pair of points in the signature space (n, m) with respect to three pairs of different relative frequencies (f_x, f_y). The indices are based on the following parameters: the offset of the second point with respect to the first point (two dimensions), the directions of the gradients of the signature at these two points (six dimensions), the amplitude ratios of the signature values at these two points (three dimensions). A hash table is used to store all the indices and the additional information vectors of the models. The additional information vectors include elements such as the angle θ_l of the kernels and the coordinates of the first point (n, m) that are used for deriving pose information.

The relative pose, which corresponds to relative shift of the first points of model indices with respect to the first points of object indices in n and m, and the angles of the kernels are derived in the process of indexing as by products. These numbers can subsequently be translated to the relative slant, tilt, scale and swing between the image patch and the model.

In Eq. (4), we see that the affine transform introduces a pose dependent shift of the signatures as well as a scalar $|C|$. Using the ratio of amplitudes as part of the multi-dimensional index eliminates the effect of this scalar and also yields invariance to multiplicative variations of image intensity.

1.4 Experimental Results

1.4.1 Recognition of Single Patch Isolated Objects

This section describes experimental results using the above mentioned approach for affine-invariant recognition. In these experiments, according to the notation of Eq. (9) in Section 1.2, the kernels $g_{mn}^{f_x,f_y,\theta_l}(x,y)$ employ a set of Standard Deviations $\{\sigma_{X_m},\sigma_{Y_n} = 8...24\}$, a set of relative frequency coefficients $\{(f_x,f_y) = (1,1),(4,4),(7,7)\}$, and 24 orientations θ_l in steps of 7.5 degrees. For a given image patch $I(x,y)$, a set of spectral signatures S^{f_x,f_y,θ_l} is generated by correlating it with the above kernels.

As elaborated in Section 1.2, these signatures are used along with a MDI scheme for affine-invariant recognition. For each model to be included in the hash table, signatures are generated using the kernels $g_{mn}^{f_x,f_y,\theta_l}(x,y)$, and the set of 11-dimensional indices are computed. Each index is included as an entry in the hash table along with the pose parameters of the model, represented by n, m and θ_l. Given an image patch to be recognized invariant to affine transformation, its signatures and 11-dimensional indices are generated in an identical fashion. These indices are then compared with indices in the hash table and each matching index adds one vote for the corresponding models pointed to by a pointer in that entry. In addition, pose information is derived as described in Section 1.3. The total number of votes accumulated by each model (with pose) over all the indices of the test image is the matching score for that model.

We use a dataset of 26 objects (displayed in Fig. 6) in our initial experiments. Since the experiments are mainly performed to test the pictorial affine invariant recognition scheme in this section, every object in the dataset is considered as a single patch. These models consist of randomly selected, real gray-level images (128×128) of objects with some amount of texture as well. A hash table is created using a single set of signatures from each model. Experiments are performed under varied conditions of slant, tilt, scale and swing and yield close to 100% correct

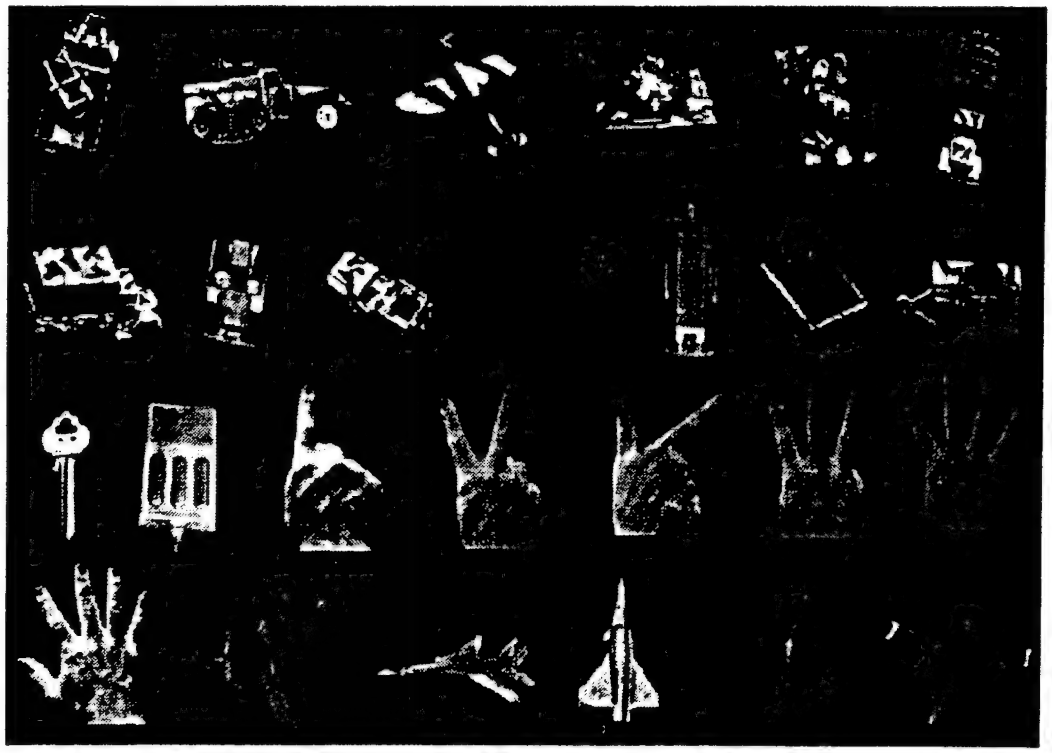


Figure 6: 26 model objects in the dataset. Close to 100% recognition is achieved over a wide range of slant, tilt, scale, and swings. Note that many of the models (such as hands) are quite similar in appearance and are still correctly classified.

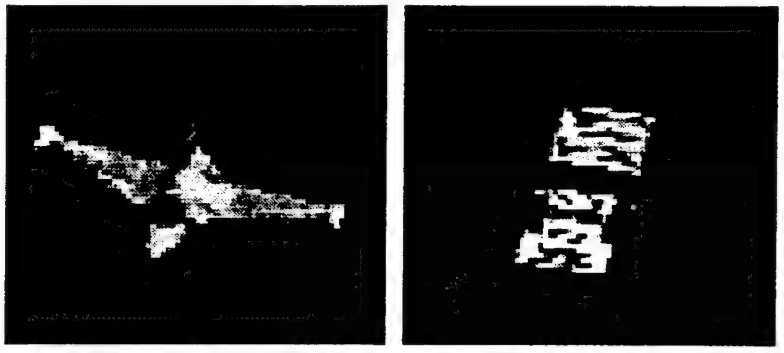


Figure 7: Two test images that correspond to affine-transformed models (compare to airplane and truck in Fig. 6).

recognition rates as illustrated in figures 8, 9, 11 and 14. In addition, the pose of each model is also estimated correctly in all experiments.

Robust recognition is achieved over a range of more than 3 octaves of scaling, slant angles of more than 80 degrees (foreshortening ratio of 1:7), and image swing and shape tilt of 360 degrees. Two of the successfully recognized test images are displayed in Fig. 7.

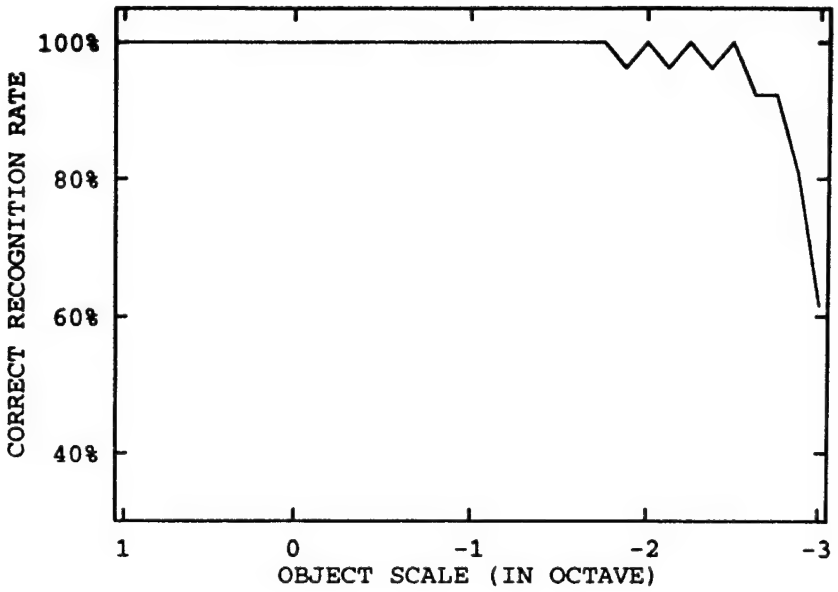


Figure 8: Correct recognition rates for scaled objects. The experiments are performed with 26 models and all the test objects are scaled versions of the models at different scales from 1 octave to -3 octave (e.g. from scale factors of 2 down to 0.125).

In recognition experiments in varied scale, as illustrated in Fig. 8, the method achieves 100% recognition rate even when images are down scaled by 2.5 octaves. It should be noted that the images scaled halfway in between the decimation interval for our Gabor kernels are still correctly recognized. The maximum error in pose estimation is 1.09 of the scale factor. In swing and tilt experiments, the recognition rates are examined for the full 360° and are found to be 100%. Due to the constant recognition rates, graphs are not presented for tilt and swing. In slant experiments, the images are foreshortened only in one direction. Minimal foreshortening factor is around 0.0743, which corresponds to a slant angle of 85.4 degrees. Fig. 9 shows the correct recognition rates for different slant angles.

Figure 10 illustrates three test images that are noisy versions of the corresponding model in Fig. 6. Experiments are carried out with additive white noise, low-frequency colored noise (normalized low pass cutoff frequency $= \pi/2$), and high-frequency colored noise (normalized high pass cutoff frequency $= \pi/2$). For each kind of noise, we experimentally find the largest noise level at which successful recognition with correct pose estimation is obtained. As seen in Fig. 10, the test image is successfully recognized in all three cases at very high noise levels, demonstrating that the scheme is quite robust to additive noise. The Gabor kernels capture the image information

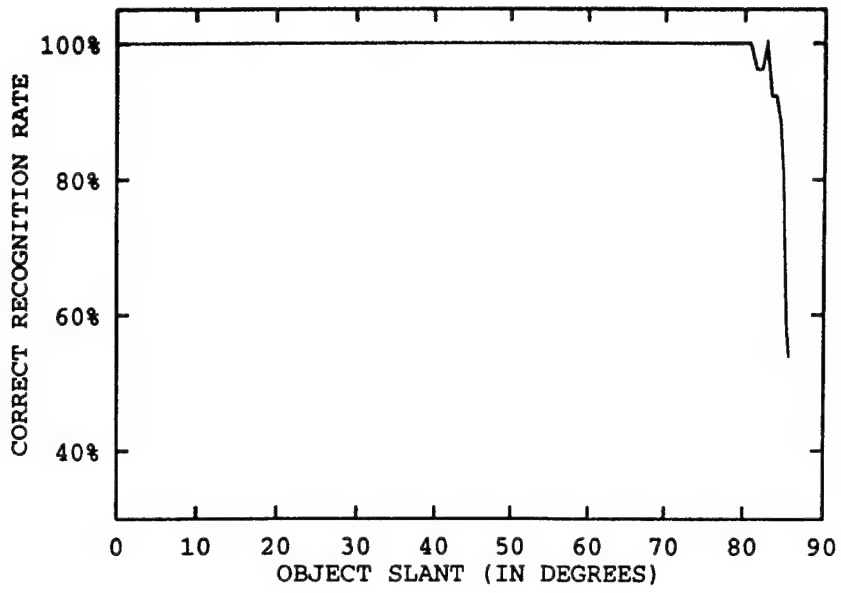
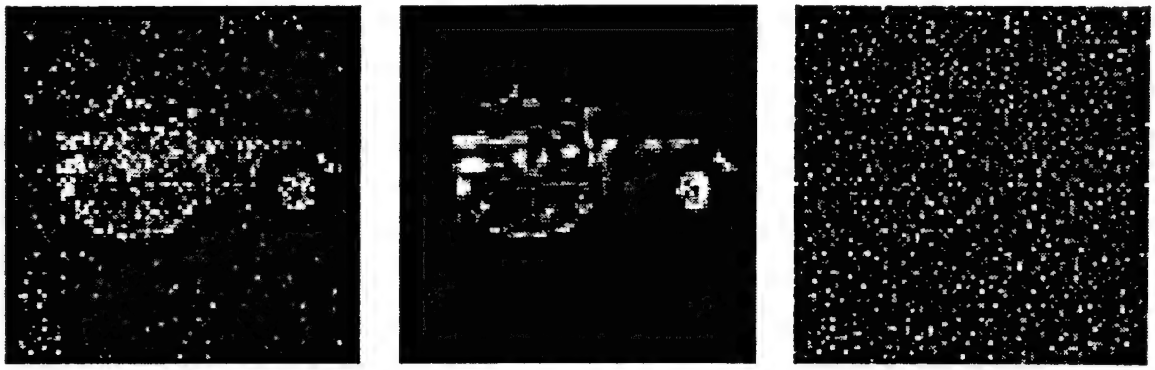


Figure 9: Correct recognition rates for slanted objects. The model dataset consists of 26 objects and all the test objects are slanted versions of the models at different slant angles from 0° to 86°.



a b c

Figure 10: (a) Successfully recognized test image with additive white noise (SNR=-1.8 dB). (b) Successfully recognized test image with additive low-frequency colored noise (SNR=5.0 dB). (c) The test image is recognized even though it is hardly seen (additive high-frequency colored noise of SNR=-17.0 dB).

mostly in the low and middle frequencies, and thus the scheme is almost insensitive to high-frequency noise (Fig. 10(c), SNR=-17 dB) since this noise is outside the frequency range of the kernels. The scheme is also quite resistant to white noise (up to SNR=-1.8 dB), and less resistant (up to SNR=5dB) to low-frequency noise for the same reason. Thus, we can conclude that the

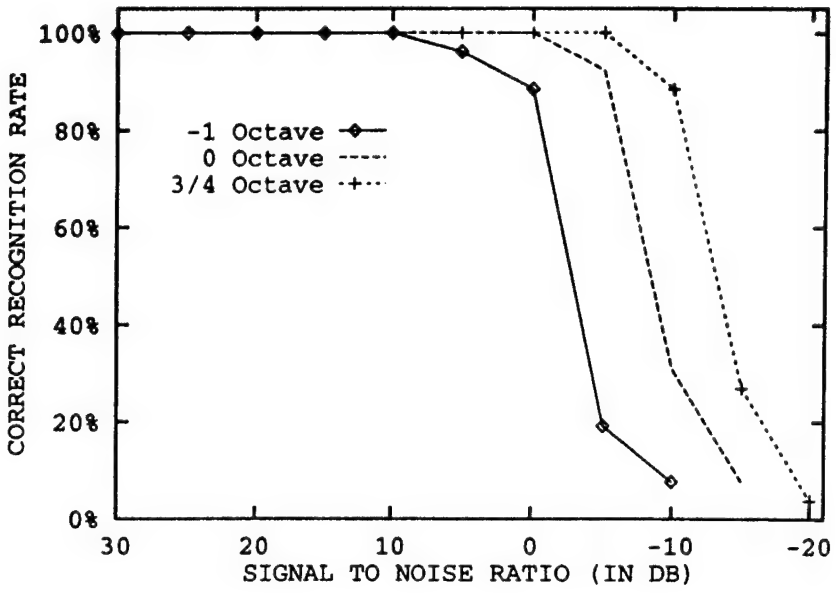
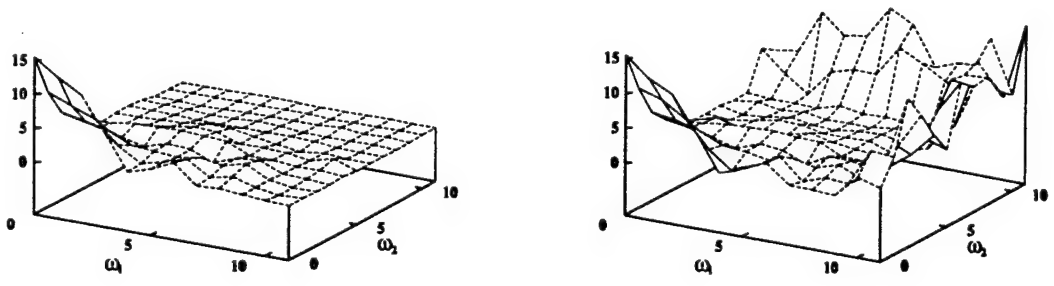


Figure 11: Correct recognition rates for white-noise corrupted objects. The model dataset consists of 26 objects and the test objects are noisy and scaled versions of the models at three different scales.

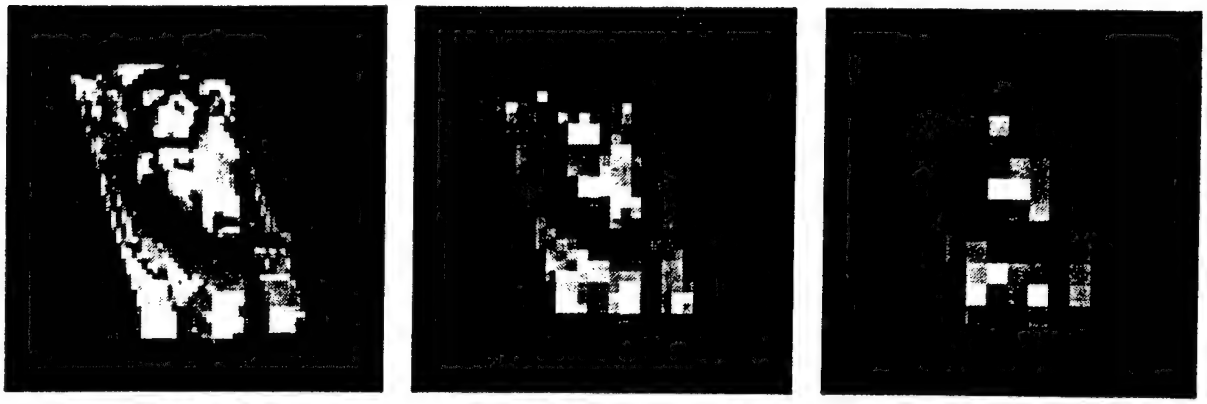
overall recognition scheme is quite robust to the effects of additive noise and clutter. Fig. 11 gives the correct recognition rates for white-noise corrupted images with different levels of SNR.



a b

Figure 12: (a) The signature of the noiseless truck model for Fig. 10. (b) The signature of the noisy image in Fig. 10c. Only the high frequency regions are affected while the low and middle frequency responses still allow for robust recognition.

In order to explain the surprisingly good recognition in high frequency noise, we demonstrate



a b c

Figure 13: (a) Original high resolution model (128×128). (b) Medium resolution test image (64×64). (c) Low resolution test image (32 × 32).

the effects of high frequency noise on the signature in Fig. 12(a) and Fig. 12(b). As observed, the noisy signature in Fig. 12(b) (which corresponds to picture in Fig. 10(c)) is affected only at only two boundaries while the rest is almost identical to the signature in Fig. 12(a). This explains why the computer is still able to recognize the image in Fig. 10(c) which looks completely noisy to the human eye.

Figures 13(b-d) illustrate test images that have reduced resolution with respect to the model image in Fig. 13(a) (which is 128 × 128 in size). Experiments were performed over all the 26 models for each of these resolutions. To simulate affine transformation in addition to the effects of reduced resolution, all the test images correspond to model images and are scaled by a factor of 0.8 and swung by 30 degrees. Over all the 26 models in the dataset, the medium resolution (64 × 64) set of test images (see Fig. 13(b)) yields 100% successful recognition with the correct pose estimated in all tests. In the low resolution (32 × 32) experiments of Fig. 13(c), 96% of the test images were successfully recognized along with accurate pose estimation. These results show that the representation and recognition scheme is quite robust to significant degradation that correspond to lower resolution. Such degradation could occur from large viewing distances. In Fig. 14, the correct recognition rates under different levels of resolution reduction are given.

1.4.2 Recognition of Multiple Patches and Non-Isolated Objects Using Model Based Segmentation

In all the experiments described in Section 1.4.1, we consider every image as a single patch. For recognition of multiple objects in one image, we first have to obtain the spectral signature for every local patch. Most objects encountered in daily life are composed of a number of primitive

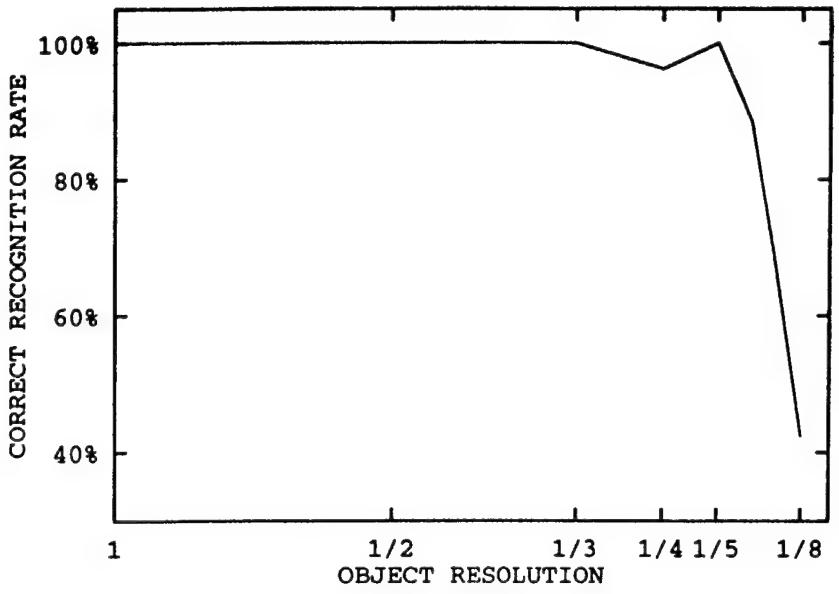
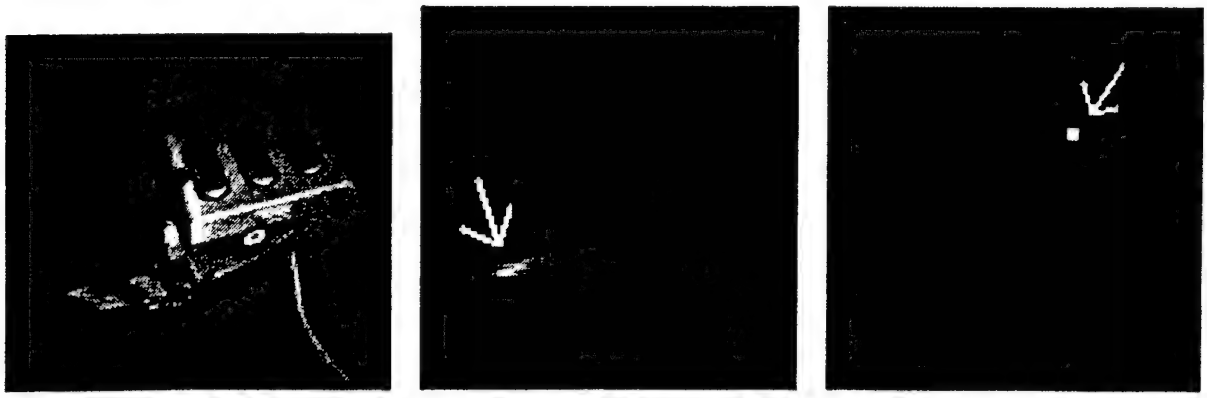


Figure 14: Correct recognition rates for objects at different resolutions. The experiments are performed with 26 object models and the test objects are derived from the models through low-pass filtering and down-sampling.

standard shapes. Hence, it's assumed that a large set of man made objects (which include most objects in our model dataset) can be represented by a small set of standard primitives. In our case, we prefer simple primitives such as rectangles, semicircles etc. that can approximate many man made flat surfaces. For detection of such primitives in the image, it is advantageous to use their boundaries (edges) since this information is more stable in varying conditions of illumination.

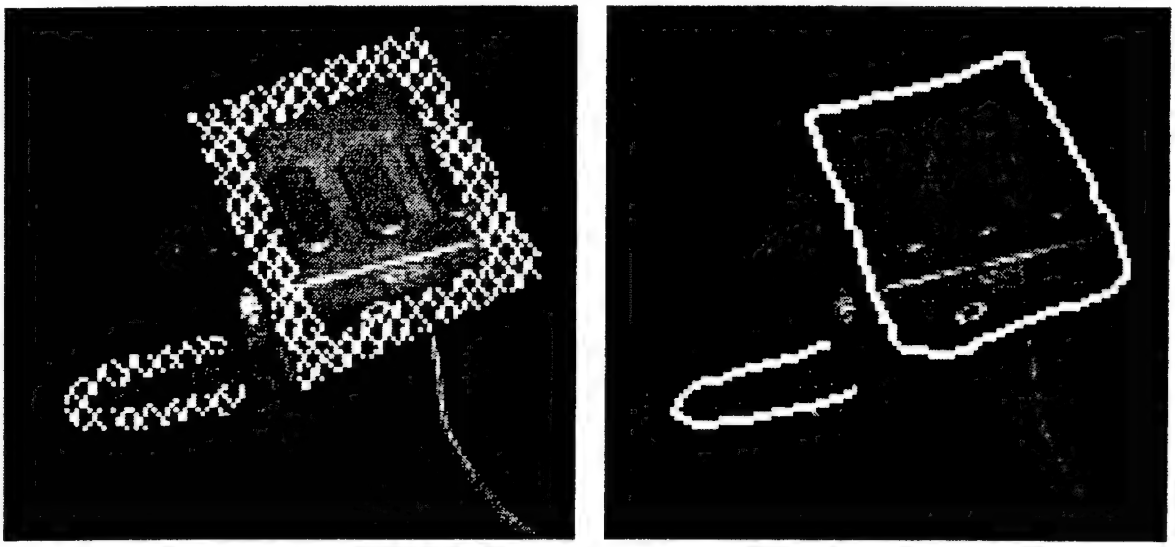
For the detection of standard primitives that may vary in their proportions, sizes and orientations, is it necessary to generate a set of boundary models. These models cover the boundaries of standard primitives with strips to allow for local variations (as in Fig. 16(a)). When an edge map of an image is convolved with the set of strip models, the peaks detected indicate possible existence of shapes similar to the corresponding model set. In order to detect primitive parts in the image with affine-invariance, the set of strip models must include all the affine-transformed versions of each primitive part. Since the number of strip models in the library might be very large, an efficient representation approach needs to be developed.

We choose to employ here the Karhunen-Loeve (K-L) transform which is commonly used as an optimal compression technique for images. For this application, the K-L transform is employed to compress strip models. A large number of strip models (approximately 44,000 templates) are



a b c

Figure 15: (a) Recognition of neighboring objects (airplane and mouse) in background clutter. (b) Convolution score of the edge map of the image with the semi-circular strip template set. (c) Convolution score of the edge map of the image with the rectangular strip template set.



a b

Figure 16: (a) Initial estimation of shapes and poses. The rectangular and semicircular strip models that correspond to peaks found in Fig. 15(b) and Fig. 15(c) are overlaid on the the image with their correct affine transforms. (b) Final results after flexible matching.

approximated by only 10 eigentemplates. The eigentemplates are then convolved with the edge map of the image. Fig. 15 shows the convolution scores of the edge map with the K-L based set of strip templates. At each point, the score denotes the highest score of the convolutions of all strip models with the edge image. As seen in Fig. 15, sharp peaks are obtained for a slanted semi-circle (Fig. 15(a)) and a rectangle (Fig. 15(b)). The peaks of the convolution provide a robust model based segmentation of the image as illustrated in Fig. 16(a). It also provides an affine invariant initial identification of the primitives in the image and their poses.

At a second stage, for more accurate segmentation, an algorithm of a flexible matching of the primitive parts to the real edges in the image is employed. The principle of elastic matching implemented is similar to the snakes [15]. But our flexible matching algorithm differs in the aspect that our primitives models are attached to a virtual elastic sheet that is distorted to match the exact shape beneath it. When the elastic sheet finally locks on to the real image, the shape of the corresponding part is determined. Fig. 16(b) shows the final result of flexible matching for the image. The primitive parts are also found the same way in the model dataset. The signatures of the segmented parts in the image are then matched against the primitive parts of the models. Since the parts are segmented and isolated, the signatures obtained are not affected by neighboring objects and the background. For example, we display the objects that match rectangular primitives in our 26 model dataset of Fig. 6 and their matching scores with the segmented mouse in Fig. 17. Even though the segmentation of the mouse in Fig. 16(b) includes small parts of the airplane wings, the matching scores of the signatures clearly classifies it correctly. Based upon the segmentation results, the airplane and the mouse are successfully recognized.

This approach is run on a Pentium Pro (200 MHz) personal computer. 28224 eleven dimensional indices are generated for each image patch. In the experiments of isolated object recognition, average recognition time is around 20 seconds. It takes around 3.5 minutes to recognize objects in cluttered images. A detailed and general analysis of time and memory requirements for multidimensional indexing can be found in [13].

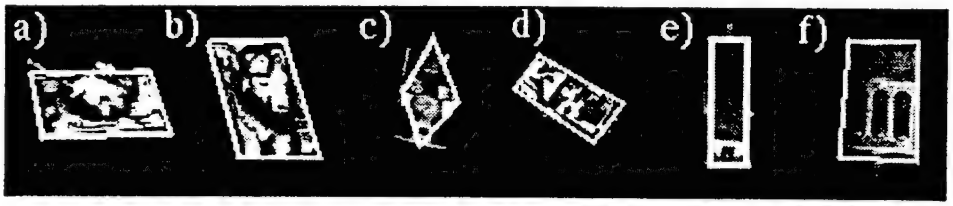


Figure 17: The models matching the rectangular primitive. The matching scores of the mouse in Fig. 16(b) to these models are: 0.53, 0.59, 0.50, 0.45, 0.50 and 0.80 for a) to f).

1.5 Conclusion

We present here an approach for affine-invariant object recognition by pictorial recognition of image patches that correspond to object surfaces that are roughly planar. Each surface can be recognized separately invariant to its 3D pose, employing novel Affine-Invariant Spectral Signatures (AISSs). The 3D-pose invariant recognition is achieved by correlating the image

with a novel configuration of Gabor kernels and extracting local spectral signatures. The local spectral signature of each image patch is then matched against a set of pictorial models using Multi-Dimensional Indexing (MDI) in the frequency domain. Affine-invariance of the signatures is achieved by a new log-log sampling configuration in the frequency domain which can be represented by short-term Fourier transform or by Gabor transform in two orthogonal axes. In our experiments, we find that spectral signatures have a significant discriminative power even without phase information. 100% correct affine-invariant recognition is obtained in a range of more than 3 octaves of scaling and slant angles of more than 80 degrees, with image swing and shape tilt of 360 degrees with a dataset of 26 gray-level models. Experiments also reveal that the method works with severe additive white and colored noise (SNR of -17 dB to 5 dB) and degraded resolution. To overcome the problem of recognition of non-isolated objects, we develop a model based segmentation scheme. This scheme enables to extract isolated signatures of image regions, which are affine projection of a set of basic geometric shapes such as rectangle, triangle, semicircle etc.

2 A Volumetric/Iconic Frequency Domain Representation for Objects with application for Pose Invariant Face Recognition

2.1 Introduction

A major problem in 3-D object recognition is the method of representation, which actually determines to a large extent, the recognition methodology and approach. The large variety of representation methods presented in the literature do not provide a direct link between the 3-D object representation and its 2-D views. These representation methods can be divided into two major categories: object centered and viewer centered (iconic). Detailed discussion are included in [25] and [21]. An object centered representation describes objects in a coordinate system attached to objects. Examples of object centered methods of representation are spatial occupancy by voxels [25, pp. 468-469], constructive solid geometry (CSG) [25, pp. 468], superquadrics [32] [9], etc. However, object **views** are not explicitly stored in such representations and therefore such datasets do not facilitate the recognition process since the images **cannot be directly indexed** into such a dataset and need to be matched to **views** generated by perspective/orthographic projections. Since the viewpoint of the given image is a priori unknown, the recognition process becomes computationally expensive. The second category i.e. viewer centered (iconic) representation is more suitable for matching a given image with such a dataset, since the dataset also is comprised of various views of the objects. Examples of viewer centered methods of representation are aspect graphs [26], quadtrees [21], Fourier descriptors [45], moments [23], etc. However, in a direct viewer centered approach, the huge number of views needed to be stored renders this approach impractical for large object datasets. Moreover, such an approach does not automatically provide a 3-D description of the object. For example, in representations by aspect graphs [26], qualitative 2-D model views are stored in a compressed graph form, but the view retrieval requires additional 3-D information in order to generate the actual images from different viewpoints. In principle, viewer centered aspect graph approaches do not offer significant advantage over object centered approaches. In summation, viewer centered and object centered representations have complementary merits that could be augmented in a merged representation - as proposed in this section.

A first step in unifying object and viewer centered approaches is provided by our recently developed Affine Invariant Spectral Signatures (AISS) approach [7] [6] [5], which is based on an iconic 2-D representation in the frequency domain. However, the AISS is fundamentally different from other viewer centered representations since each 2-D shape representation encapsulates **all**

the appearances of that shape from any spatial pose. It also means that the AISS enables to recognize surfaces which are approximately planar, invariant to their pose in space. Although this approach is basically viewer centered, it has the advantage of directly linking 3-D model information with image information, thus merging object and viewer centered approaches. Hence, to generalize the AISS it is necessary to extend it from 2-D or flat shapes to general 3-D shapes. Towards this end, we describe in Section 2.2, a novel representation of 3-D objects by their 3-D spectral signatures which also captures all the 2-D views of the object and therefore facilitates direct indexing of a given image into such a dataset.

As a demonstration of the V/ISS representation, it is applied for estimating pose of faces and face recognition in Section 2.3. Range image data of a human head is used to construct the V/ISS model of a simulated "generic" face. We demonstrate that reconstruction from slices of the V/ISS results are accurate enough to recognize faces from different spatial poses and scales. In Section 2.3, we describe the matching technique by means of which a gray scale image of a face is directly indexed into the 3-D V/ISS model based on fast matching by correlation in a 4 dimensional Fourier space. In our experiments (described in Section 2.5), we demonstrate how the range data generated from a model is used to estimate the pose of a person's face in various images. We also demonstrate the robustness of our 2-D slice matching process by recognizing faces with different poses from a dataset of 40 subjects, and present statistics of the matching experiments.

2.2 Volumetric/Iconic Spectral Signature

In this section, we describe a novel formulation that merges the 3-D object centered representation in the frequency domain to a continuum of its views. The views are also expressed in the frequency domain. The following formulation describes the basic idea.

Given an object O, which is defined by its spatial occupancy on a discrete 3-D grid as a set of voxels $\{V(x,y,z)\}$, we assume without loss of generality, that the object is of equal density. Thus, $V(x,y,z) = 1 \ \ \forall \ \ \{x,y,z\} \in O$ and $V(x,y,z) = 0$ otherwise. The 3-D Discrete Fourier Transform (DFT) of the object is given by

$$\mathcal{V}(u,v,w) = \mathcal{F}\{V(x,y,z)\} = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1}\sum_{w=0}^{N-1} V(x,y,z)e^{-\jmath\frac{2\pi}{N}(ux+vy+wz)} \tag{11}$$

where $\jmath = \sqrt{-1}$. The surface of the object is derived from the gradient vector field

$$\nabla V(x,y,z) = [\boldsymbol{k}_x\frac{\partial}{\partial x} + \boldsymbol{k}_y\frac{\partial}{\partial y} + \boldsymbol{k}_z\frac{\partial}{\partial z}]V(x,y,z) \ \ , \tag{12}$$

where $\boldsymbol{k_x}$, $\boldsymbol{k_y}$ and $\boldsymbol{k_z}$ are the unit vectors along the x, y and z axes. The 3-D Discrete Fourier Transform (DFT) of the surface gradient is given by the frequency domain vector field:

$$\mathcal{F}\{\boldsymbol{D}(x,y,z)\} = \jmath\frac{2\pi}{N}(\boldsymbol{k_x}u + \boldsymbol{k_y}v + \boldsymbol{k_z}w)\mathcal{V}(u,v,w)\,. \tag{13}$$

Let the object be illuminated by a distant light source[3] with uniform intensity Υ and direction $\boldsymbol{i} = i_x\boldsymbol{k_x} + i_y\boldsymbol{k_y} + i_z\boldsymbol{k_z}$. We assume the surface has an albedo $A(x,y,z)$. For a voxel based description, the gradient magnitude $\mid \nabla V \mid \approx K$ (constant). ∇V may be estimated as $V * \nabla G$. Thus the surface normal is the given by $\frac{\nabla V}{K}$. We assume that O has a Lambertian surface with constant albedo. Thus points on its surface have a brightness proportional to

$$\begin{aligned} B_{\boldsymbol{i}}(x,y,z) &= B^+_{\boldsymbol{i}}(x,y,z) + B^-_{\boldsymbol{i}}(x,y,z) \\ &= \frac{\Upsilon A}{K}[i_x\frac{\partial}{\partial x}+i_y\frac{\partial}{\partial y}+i_z\frac{\partial}{\partial z}]V(x,y,z) \end{aligned} \tag{14}$$

where $B^+_{\boldsymbol{i}}$ and $B^-_{\boldsymbol{i}}$ are the positive and negative parts. The function $B^-_{\boldsymbol{i}}(x,y,z)$ is not a physically realizable brightness and is introduced only for completeness of Eq. (14). The separation of the brightness function into positive and negative components is used to consider only positive illuminations. The negative components are disregarded in further processing, as this function is separable only in the spatial domain. As elaborated in Section 2.2.1, $\boldsymbol{B_i}^-$ can be eliminated using a local Gabor transform.

It is also necessary to consider the viewing direction when generating views from the V/ISS. The brightness function $B_{\boldsymbol{i}}(x,y,z)$ is decomposed as a 3-D vector field by projecting onto the surface normal at each point of the surface. This enables the correct projection of the surface from a given viewpoint. As noted earlier, the surface normal is given by $\frac{\nabla V}{K}$. Thus the new vectorial brightness function $\boldsymbol{B_i}$ is given by

$$\boldsymbol{B_i}(x,y,z) = \frac{\Upsilon A}{K}\left[\boldsymbol{i}\cdot\nabla V(x,y,z)\right]\frac{1}{K}\nabla V(x,y,z)\,. \tag{15}$$

The 3-D Fourier transform of this model is a complex 3-D vector field $\boldsymbol{\mathcal{V}_i}(u,v,w)$=$\mathcal{F}\{\boldsymbol{B_i}(x,y,z)\}$. The transform is evaluated as:

$$\boldsymbol{\mathcal{V}_i}(u,v,w) = \frac{\Upsilon A}{K}\jmath\frac{2\pi}{N}(i_xu + i_yv + i_zw)\mathcal{V}(u,v,w) * \frac{1}{K}\jmath\frac{2\pi}{N}(\boldsymbol{k_x}u + \boldsymbol{k_y}v + \boldsymbol{k_z}w)\mathcal{V}(u,v,w) \tag{16}$$

where $*$ denotes convolution. Variation in illumination only emphasizes the amplitude of $\boldsymbol{\mathcal{V}_i}$ in the (i_x, i_y, i_z) direction, but does not change its basic structure. The absolute value of $\boldsymbol{\mathcal{V}_i}(u,v,w)$ is defined as the Volumetric/Iconic Spectral Signature (V/ISS).

[3]Additional light sources can be easily handled using superposition.

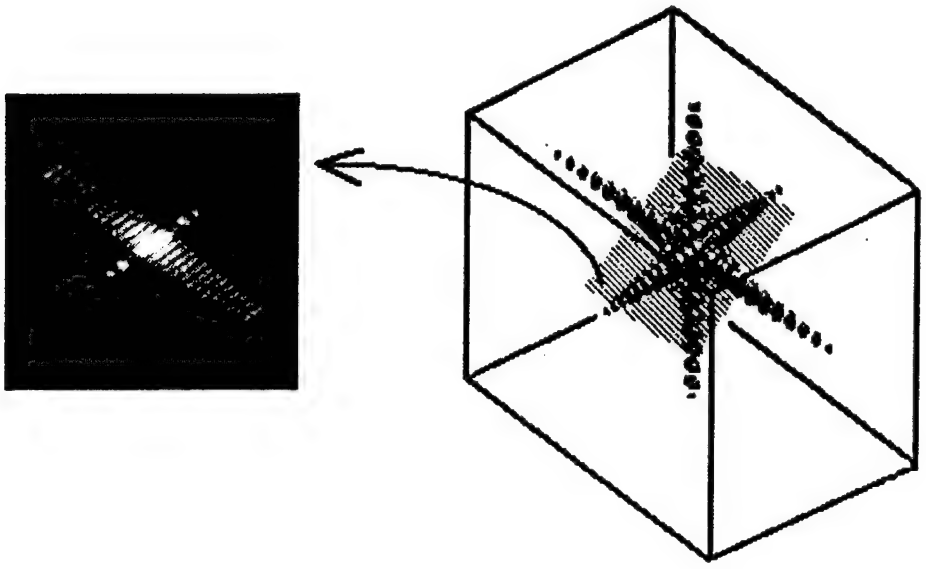


Figure 18: The Projection-Slice Theorem: A slice of the 3-D Fourier Transform of a rectangular block (on the right) is equivalent to the 2-D Fourier Transform of the projection of the image of that block (on the left).

2.2.1 Projection Slice Theorem and 2-D Views

The function $\mathcal{V}_i(u,v,w)$ is easily obtained, given the object O. To generate the view of the object, we resort to 3-D extensions of the Projection-Slice Theorem [24] [36] that projects the 3-D vector field $\mathcal{V}_i(u,v,w)$ onto the central slice plane normal to the viewpoint direction. Fig. 18 illustrates the principle by showing the slice derived from the 3-D DFT of a rectangular block. Orthographically viewing the object from a direction $\boldsymbol{c} = c_x\boldsymbol{k}_x + c_y\boldsymbol{k}_y + c_z\boldsymbol{k}_z$, results in an image $I_{\boldsymbol{C}}(x',y')$, which has a 2-D DFT given by $\mathcal{I}_{\boldsymbol{C}}(u',v')$. To find $I_{\boldsymbol{C}}(x',y')$ and its DFT $\mathcal{I}_{\boldsymbol{C}}(u',v')$, it is necessary to project the vector brightness function $\boldsymbol{B}_i^+(x,y,z)$ along the viewing direction $\boldsymbol{c}$ after removing all the occluded parts from that viewpoint. The vectorial decomposition of the brightness function along the surface normals as given by Eq. (15) compensates for the integration effects of projections of slanted surfaces. This explains the necessity of using a **vectorial** frequency domain representation.

Removing the occluded surfaces is not a simple task if the object O is not convex or if the scene includes other objects that may partially occlude O. For now, we shall assume that O is convex and is entirely visible. This assumption is quite valid for local image analysis where a local patch can always be regarded as either entirely occluded or visible. Also, for local analysis $B_{\boldsymbol{i}}^-(x,y,z)$ is not a major problem. The visible part of $B_{\boldsymbol{i}}(x,y,z)$ from direction $\boldsymbol{c}$, denoted by

$B_{ic}(x,y,z)$, is given by

$$B_{ic}(x,y,z) = \frac{\Upsilon A}{K}\mathbf{hwr}\left[i\cdot\nabla V(x,y,z)\right]\frac{1}{K}\mathbf{hwr}\left[c\cdot\nabla V(x,y,z)\right]. \tag{17}$$

where $\mathbf{hwr}[\alpha]$ is the "half wave rectified" value of α, i.e. $\mathbf{hwr}[\alpha] = \alpha$ if $\alpha \geq 0$ and $\mathbf{hwr}[\alpha] = 0$ if $\alpha < 0$.

Now $\mathcal{V}_{ic}(u,v,w)$ can be obtained from $B_{ic}(x,y,z)$ simply by calculating the DFT,

$$\mathcal{V}_{ic}(u,v,w) = \mathcal{F}\{B_{ic}(x,y,z)\} \ . \tag{18}$$

The image DFT $\mathcal{I}_c(u',v')$ is obtained using the Projection-Slice Theorem [24] [36] by slicing $\mathcal{V}_{ic}(u,v,w)$ through the origin $u=v=w=0$ with a plane normal to $\boldsymbol{c}$, i.e. $uc_x + vc_y + wc_z = 0$. $\mathcal{I}_c(u',v')$ is derived by sampling $\mathcal{V}_{ic}(u,v,w)$ on this plane. An example of such a slicing operation is illustrated in Fig. 18. Note that $\mathcal{V}_{ic}$ actually encapsulates both the objects 3-D representation and the continuum of its view-signatures, which are stored as planar sections of $|\ \mathcal{V}_{ic}\ |$. As we see from Eq. (16), variations in illumination only emphasizes the amplitude of $\mathcal{V}_i$ in (i_x,i_y,i_z) direction, but do not change its basic structure. Thus, it is feasible to recognize objects that are illuminated from various directions by local signature matching methods as described in Section 2.2.3, while employing the same signature.

2.2.2 Local Signature Analysis in 3-D

Local signature analysis is implemented by windowing $\boldsymbol{B_i}$ with a 3-D Gaussian centered at location (μ_x,μ_y,μ_z) and proceeding as in Eq. (15) on the windowed object gradient. Such local frequency analysis removes the self-occluded parts. Therefore, we use in our frequency analysis and representation, the Gabor Transform (GT) instead of the DFT. The transition required from the DFT to the GT is quite straightforward. The object O is windowed with a 3-D Gaussian to give

$$\boldsymbol{B_{iG}} = \mathcal{G}[\boldsymbol{B_i}] = \boldsymbol{B_i} e^{\{-\frac{1}{2}[(\frac{x-\mu_x}{\sigma_x})^2+(\frac{y-\mu_y}{\sigma_y})^2+(\frac{z-\mu_z}{\sigma_z})^2]\}} \ . \tag{19}$$

The equivalent **local V/ISS** is given by

$$\boldsymbol{\mathcal{V}_{iG}}(u,v,w) = \boldsymbol{\mathcal{V}_i}(u,v,w) * \left[e^{-(\frac{2\pi}{N})^2\frac{1}{2}[(u\sigma_x)^2+(v\sigma_y)^2+(w\sigma_z)^2]} \cdot e^{j\frac{2\pi}{N}[u\mu_x+v\mu_y+w\mu_z]}\right]. \tag{20}$$

The important outcome from this are: 1) The Radon transform and the Projection-Slice Theorem [24], [36] can be still employed for local space-frequency signatures of object parts. 2) In local space-frequency analysis, $\boldsymbol{B_i}$ almost always does not contain a problematic B_i^- part, which can be eliminated by the windowing function. We note that for most local surfaces, $[\boldsymbol{B_i}\cdot\boldsymbol{c}] \simeq B_{ic}$, as the local analysis approximates the $\mathbf{hwr}[\cdot]$ function with respect to viewing direction c. Hence, the V/ISS of B_{ic} is a general representation of a local surface patch of $V(x,y,z)$.

2.2.3 Indexing using V/ISS

As explained in Section 2.2.1, the V/ISS is a continuum of the 2-D DFT of views of the model. To facilitate indexing into the V/ISS data structure, we consider the V/ISS slice plane $uc_x + vc_y + wc_z = 0$, where $[c_x, c_y, c_z]^T$ are the direction cosines of the slice plane normal. We define a 4-D pose space in the frequency domain which consists of the azimuth α and elevation ϵ, defining the slice plane normal with respect to the original axes, the in-plane rotation θ of the slice plane and the scale ρ which changes with the distance to the viewed object. Fig. 20 illustrates the coordinate system used. $[c_x, c_y, c_z]^T$ are related to the azimuth α and elevation ϵ as follows

$$\begin{bmatrix} c_x \\ c_y \\ c_z \end{bmatrix} = \begin{bmatrix} \cos\alpha\cos\epsilon \\ \sin\alpha\cos\epsilon \\ \sin\epsilon \end{bmatrix} \quad \begin{matrix} -\pi/2 \le \alpha \le \pi/2 \\ -\pi/2 \le \epsilon \le \pi/2 \end{matrix} \tag{21}$$

We note again that slices of the V/ISS are planes which are parallel to the imaging plane. Thus the image plane normal and the slice plane normal coincide. By using 3-D coordinate transformations (see Fig. 20) we can transform the frequency domain V/ISS model to the 4-D pose space $(\alpha, \epsilon, \theta, \rho)$. Let (u, v, w) represent the original V/ISS coordinate system and $(\hat{u}, \hat{v}, \hat{w})$ be the coordinate system defined by the slice plane. The slice plane is within the 2-D coordinate system $(\hat{u}, \hat{v})$, where $\hat{w}$ is the normal to the slice plane (and also the viewing direction). The relation between these two systems is given by

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \cos\alpha\sin\epsilon & -\sin\alpha & \cos\alpha\cos\epsilon \\ \sin\alpha\sin\epsilon & \cos\alpha & \sin\alpha\cos\epsilon \\ -\sin\epsilon & 0 & \sin\epsilon \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{v} \\ \hat{w} \end{bmatrix}. \tag{22}$$

V/ISS slices, being 2-D DFT's of model views are further transformed to polar coordinates by considering the in plane rotation θ (equivalent to the image swing or rotation about the optical axis), and the radial frequency r_f.

$$\begin{aligned} \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} &= r_f \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \quad -\pi/2 \le \theta \le \pi/2 \\ r_f &= \sqrt{\hat{u}^2 + \hat{v}^2} \quad r_0 \le r \le r_{max} \end{aligned} \tag{23}$$

The radial frequency r_f is transformed logarithmically to attain exponential variation of r_f. Thus

$$\rho = \log_a \frac{r_f}{r_0} \tag{24}$$

The full transformation of the coordinate system to the 4-D pose space is given by

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = r_0 a^\rho \begin{bmatrix} \cos\theta\cos\alpha\sin\epsilon - \sin\theta\sin\alpha \\ \cos\theta\sin\alpha\sin\epsilon - \sin\theta\cos\alpha \\ -\sin\theta\cos\epsilon \end{bmatrix}. \tag{25}$$

Thus the 4-tuple $(\alpha, \epsilon, \theta, \rho)$ defines all the points in the 3-D V/ISS frequency space (u, v, w). We observe that the space defined by the 4-tuple $(\alpha, \epsilon, \theta, \rho)$ is redundant in the sense that infinite number of 4-tuples $(\alpha, \epsilon, \theta, \rho)$ may represent the same (u, v, w) point. However, this representation has the important advantage that every (α, ϵ) pair defines a planar slice in $\mathcal{V}_{ic}(u, v, w)$. Moreover, every θ defines an image swing and every ρ defines another scale. Thus the $(\alpha, \epsilon, \theta, \rho)$ representation significantly simplifies the indexing search for the viewing poses and scales. Now, the indexing can be simply implemented by correlation in the frequency domain to immediately determine all pose parameters by linear shifts in $(\alpha, \epsilon, \theta, \rho)$ space. The significance of this transformation to the 4-D pose space is in using the following properties. The polar coordinate transformation within the slice allows rotated image views to have 2-D frequency domain signatures which shift along the θ axis. Similarly the exponential sampling of the radial frequency r_f results is scale changes causing linear shifts along the ρ axis. Thus the new coordinate system given by $(\alpha, \epsilon, \theta, \rho)$ results in a 2-D frequency domain signature which is invariant to view point and scale and results only in linear shifts in the 4-D pose space so defined. A particular slice corresponding to a particular viewpoint is easily indexed into the transformed V/ISS by using correlation.

2.3 Pose Estimation and Recognition of Human Faces

Recognition of human faces is a hard problem for machine vision, primarily due to the complexity of the shape of a human face. The change in the observed view caused by variation in facial pose is a continuum which needs large numbers of stored models for every face. Since the representation of such a continuum of 3-D views is well addressed by our V/ISS representation, we present here, the application of our V/ISS model for pose-invariant recognition of human faces. First we discuss some of the existing work in face recognition in Section 2.3.1 followed by our approach to the problem in Section 2.3.2. We present our results in face pose estimation (Section 2.4) and face recognition (Section 2.3) and compare our results in face recognition to some other recent works using the same database [31].

2.3.1 Face Recognition: A Literature Survey

Recent works in face recognition have used a variety of representations including parameterized models like deformable templates of individual facial features [44] [38] [16], 2-D pictorial or iconic models using multiple views [12] [10], matching in eigenspaces of faces or facial features [33] and using intensity based low level interest operators in pictures. Recent significant works in face recognition have used convolutional neural networks [29] as well as other neural network

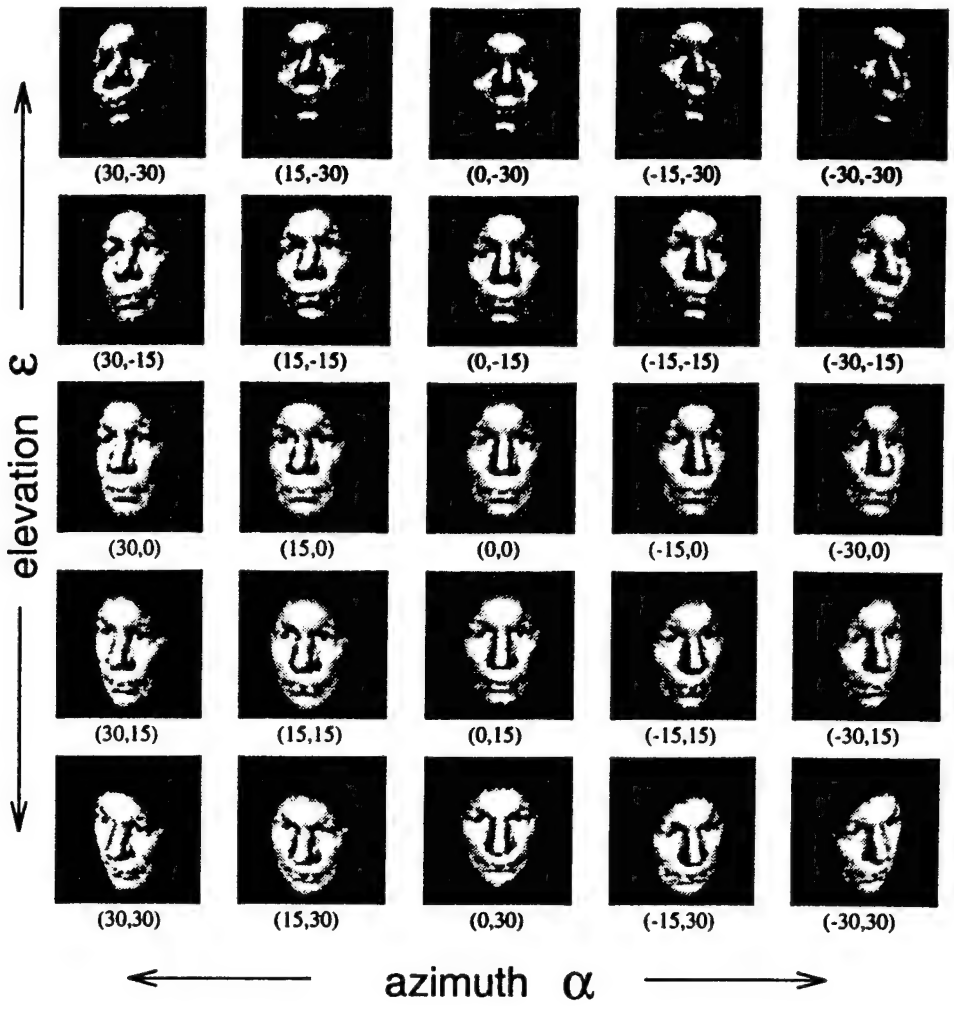


Figure 19: Reconstructions of a model face from slices of the V/ISS are shown for various azimuths and elevations. Note that all facial features are accurately reconstructed indicating the robustness of the V/ISS model.

approaches like [18] and [42]. Hidden Markov Models [37], modeling faces as deformable intensity surfaces [30], and elastic graph matching [27] have also been developed for face recognition.

Parameterized models approaches like that of Yuille *et al.* [44], use deformable template models which are fit to preprocessed images by minimizing an energy functional, while Terzopoulos and Waters [38] used active contour models of facial features. Craw *et al.* [16] and others have used global head models from various smaller features. Usually deformable models are constructed from parameterized curves that outline subfeatures such as the iris or a lip. An energy functional is defined that attracts portions of the models to pre-processed versions of the image and model fitting is performed by minimizing the functional. These models are used to track faces or facial features in image sequences. A variation is the deformable intensity surface

model proposed by Nastar and Pentland [30]. The intensity is defined as a deformable thin plate with a strain energy which is allowed to deform and match varying poses for face recognition. A 97% recognition rate is reported for a database with 200 test images.

Template based models have been used by Brunelli and Poggio [12]. Usually they operate by direct correlation of image segments and and are effective only under invariant conditions of scale orientations and illumination. Brunelli and Poggio computed a set of geometrical features such as nose width and length, mouth position and chin shape. They report 90% recognition rate on a database of 47 people. Similar geometrical considerations like symmetry [35] have also been used. A more recent approach by Beymer [10] uses multiple views and a face feature finder for recognition under varying pose. An affine transformation and image warping is used to remove distortion and bring correspondence between test images and model views. Beymer reports a recognition rate of 98% of a database of 62 people, while using 15 modeling views for each face.

Among the more well known approaches has been the eigenfaces approach [33]. The principal components of the database of normalized face images is used for recognition. The results report a 95% recognition rate of 200 faces from a database of 3000) However, variation in face pose is limited. More recent reports on a fully automated approach with extensive preprocessing on the FERET database indicate only 1 mistake on a database of 150 frontal views.

Elastic graph matching using the dynamic link architecture [27] was used quite successfully for distortion invariant recognition. Objects are represented as sparse graphs with vertices labels with multi-resolution spectral descriptions and graph edges associated with geometrical distances form the database. A recognition rate of 97.3% is reported for a database of 300 people.

Neural network approaches have also been popular. Principal components generating using an autoassociative network have been used [18] and classified using a multilayered perceptron. The database consists of 20 people with no variation in face pose or illumination. Weng and Huang used a hierarchical neural network [42] on a database of 10 subjects. A more recent approach uses a hybrid approach using self organizing map for dimensionality reduction and a convolutional neural networks for hierarchical extraction of successively larger features for classification [29]. The reported results show a 3.8% error rate on the ORL database using 5 training images per person.

In [37], a HMM-based approach is used on the ORL database. Error rates of 13% were reported using a top-down HMM. An extension using a pseudo two-dimensional HMM reduces the error to 5% on the ORL database. 5 training and 5 test images were used for each of 40 people under various pose and illumination conditions.

2.3.2 V/ISS model of faces

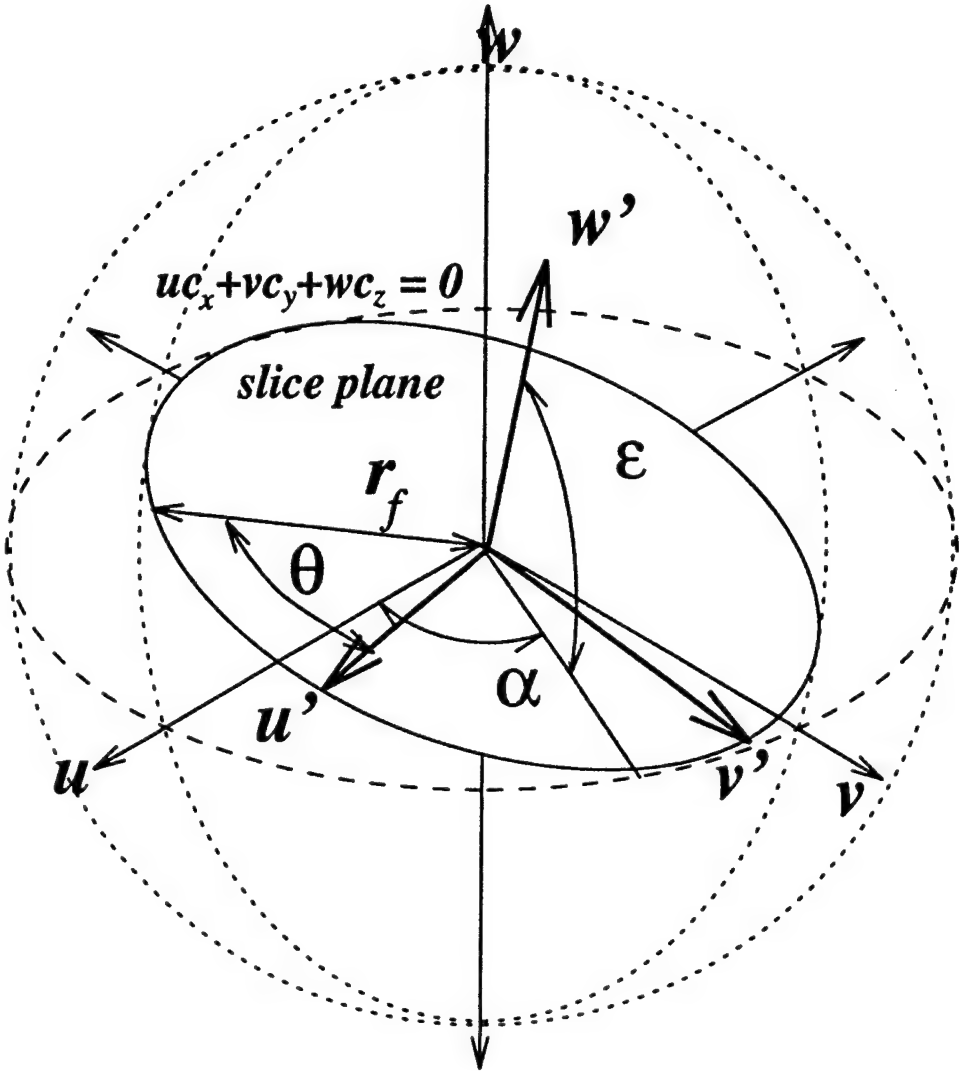


Figure 20: The frequency domain coordinate system in which the slice plane is defined. (c_x, c_y, c_z) are the direction cosines of the slice plane normal, which has an azimuth α and an elevation ϵ. Image swing is equivalent to in plane rotation θ, and viewing distance results is variation in the radial frequency r_f of the V/ISS function.

In our V/ISS model, we present a novel representation using dense 3-D data to represent a continuum of views of the face. As indicated by Eq. (18) in Section 2.2, the V/ISS model encapsulates the information in the 3-D Fourier domain. This has the advantage of 3-D translation invariance with respect to location in the image coupled with faster indexing to a view/pose of the face using frequency domain scale and rotation invariant techniques. Hence, complete 3-D pose invariant recognition can be implemented on the V/ISS.

Range data of the head is acquired using a Cyberware range scanner. The data consists of 256 × 512 range information from the central axis of the scanned volume. 360° of longitude is sampled in 512 columns and heights in the range of 25 to 35 cm is sampled in 256 rows. The data is of the heads of subjects looking straight ahead at 0° azimuth and 0° latitude corresponding to the x-axis. This model is then illuminated with numerous sources of uniform illumination thus approximating diffuse illumination in a well-lit room. The resulting intensity data in converted from the cylindrical coordinates of the scanner to Cartesian coordinates and inserted in a 3-D surface representation of the head surface as given by Eq. (14).

The facial region of interest to us is primarily the frontal region consisting of the eyes, lips and nose. A region corresponding to this area is extracted by windowing the volumetric surface model with a 3-D ellipsoid with a Gaussian fall off centered at the nose. The parameters of the 3-D volumetric mask are adjusted to ensure that the eyes, nose and lips are contained within it, with the fall off beyond the facial region. The model thus formed is a complex surface which consists of visible parts of the face from an continuous range of view centered around the x-axis or the $(0°, 0°)$ direction. The resulting model then corresponds to Eq. (17) in our V/ISS model. Applying Eq. (18), the V/ISS of the face is obtained. The V/ISS model is then resampled into the 4-D pose space using Eq. (25) as described in Section 2.2.3. Reconstructions of a range of viewpoints from a model head, from the V/ISS slices are shown in Fig. 19. We see from the reconstructions, that all relevant facial characteristics are retained thus justifying our use of the vectorial V/ISS model. This model is used in the face pose estimation experiments.

2.3.3 Indexing images into the V/ISS

Images of human faces are masked with an ellipse with Gaussian fall-off to eliminate background textures. The resulting image shows the face with the eyes nose and lips. The magnitudes of Fourier transform of the windowed 2-D face images are calculated. The windowing has the effect of focusing on local frequency components (or foveating) on the face. while retaining the frequency components due to facial features. The Fourier magnitude spectrum make the spectral signature translation invariant in the 2-D imaging plane. The spectrum is then sampled in the log-polar scheme similar to the slices of the V/ISS. As most illumination effects are typically lower frequency, band pass filtering is used to compensate for illumination.

The spectral signatures from the gray scale images are localized (windowed) log-polar sampled Fourier magnitude spectra. The continuum of slices of the V/ISS provide all facial poses, and band-passed Fourier magnitude spectrum provides 2-D translation invariant (in the imaging plane) signatures. Log-polar sampling of the 2-D Fourier spectrum allows for scale invariance

Table 1: Pose estimation errors for faces with known pose. Note these are the averaged absolute errors for angles and standard deviation of the ratio of estimated size to true size for scale.

Azimuth Error	Elevation Error	Rotation Error	Scale Std. Dev.
4.05°	5.63°	2.68°	0.0856

(translation normal to the imaging plane) and rotation invariance (within the imaging plane). This is because a scaled image manifests itself in Fourier spectrum inversely proportional to the scale and a rotated image has a rotated spectrum. Thus scaled and rotated images have signatures which are only linearly shifted in the log-polar sampled frequency domain.

The pose of a given image is determined by correlating the intensity image signature with the V/ISS in the 4-D pose space. The matching process is based on indexing through the sampled V/ISS slices and maximizing the correlation coefficient for all the 4 pose parameters. The correlation is performed on the signature gradient which reduces dependence of actual spectral magnitudes and considers only the shape of the spectral envelope. The results take the form of scale and rotation estimate along with a matching score from 0 to 1.
Similar approaches have been very sucessfully used to match Affine Invariant Spectral Signatures (AISS) [1] [3] [7] [6] [5]. References [1] and [3] already include detailed noise analysis with white and colored noise which shows robustness to noise levels of up to 0 dB SNR.

2.4 Face Pose Estimation

To verify the accuracy of the pose estimation procedure, the method is first tested on images generated from the 3-D face model. 20 images of the face in Fig. 19 are generated using random viewpoints and scales from uniform distributions. The azimuth and elevation are in the range $[-30°, 30°]$, the rotation angle is in the range $[-45°, 45°]$ and the scale in the range $[0.5, 1.5]$. These are indexed in the V/ISS pose space. The results are summarized in Table 1. An example of the correlation peak for the estimated pose in azimuth and elevation is shown in Fig. 2.4 for the test image in Fig. 2.4. The corresponding reconstructed face from the V/ISS slice is shown in Fig. 2.4.

In addition, we also show the results of pose estimation of face images of the subject with unknown pose and illumination in Fig. 24.

Figure 21: A test image with pose parameters $(14°, -8°, 41°, 1.4)$.

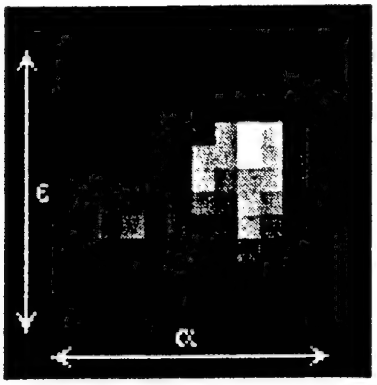


Figure 22: The correlation maximum in the azimuth-elevation dimensions of the pose space. The peak is quite discriminative as seen by relative brightness.

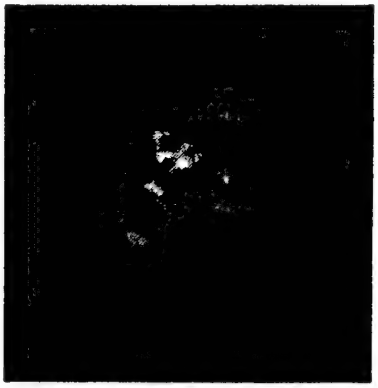


Figure 23: The reconstructed image from the slice which maximizes the correlation. Pose parameters $(10°, -10°, 40°, 1.414)$.

Table 2: Face recognition using the ORL database. Recognition rates are given for 5, 6, 7 and 8 images as V/ISS slices.

Number of Slices	5	6	7	8
Recognition Rate	92.5%	95.6%	96.6%	100%

2.5 Face Recognition Results

In this section, we describe experiments on face recognition based on the V/ISS model. The ORL database [31] is used. The ORL database consists of 10 images of each of 40 people taken in varying pose and illumination. Thus there are a total of 400 images in the database.

We select a number of these images varying from 5 to 8 as model images and the remaining images form the test set. The model images are windowed with an ellipse with a Gaussian fall-off. The recognition is robust to the window parameters selected, provided the value of σ for the Gaussian fall-off is relatively large. The images are 112×92 pixels. The window parameters chosen were 30 pixels for the longer elliptical axis aligned vertically and 22 pixels for the shorter axis aligned horizontally and $\sigma = 15$ pixels. Each window is centered at (60,46). This allows for faster processing rather than manually fitting windows to each face image. Thus, the same elliptical Gaussian window was used on all model and test images even though its axes does not align accurately with the axes of all the faces. The windowed images are transformed to the Fourier domain and then sampled in a log-polar format, now correspond to slices in a 4-D V/ISS pose space. The test images are then indexed into the dataset of slices for each person.

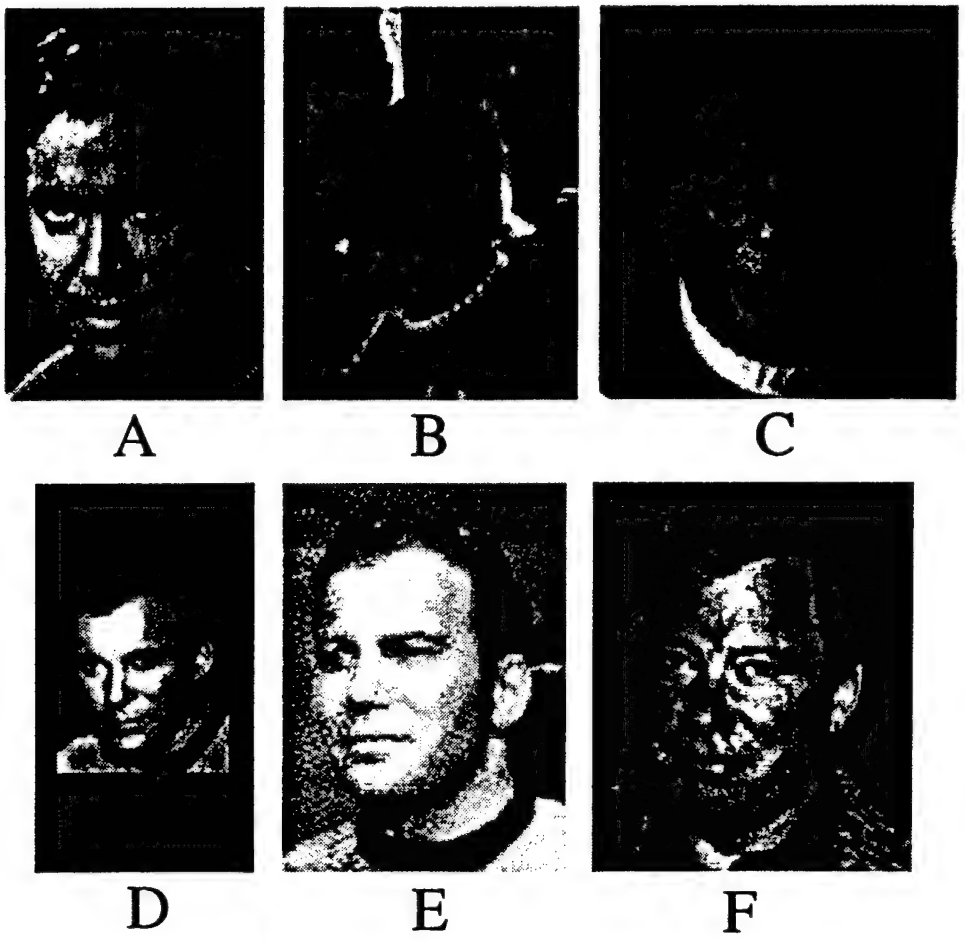


Figure 24: Using the V/ISS model, the pose of the face in the above images is estimated and the faces are recognized. The estimated poses are given in terms of the 4-tuple azimuth α, the elevation ϵ, the relative swing (rotation) θ, and the relative scale $r_0 * a^\rho$. The results are A:$(+15°, +20°, +8°, 1.6818)$, B:$(+10°, -10°, +4°, 1.0)$, C:$(+0°, -5°, -4°, 1.834)$, D:$(+15°, +25°, +4°, 1.0)$, E:$(+20°, -5°, 0°, 1.414)$ and F:$(+15°, +0°, -4°, 1.6818)$.

The recognition rates using 5, 6, 7 and 8 model images are summarized in Table 2. As can be seen, a recognition rate of 92.5% is achieved when using 5 slices. This increases to 100% when using 8 slices in the model. A few of the test images that are recognized are shown in Fig. 25. Computationally each face indexing takes about 320 seconds when using 5 slices and up to about 512 seconds when using 8 slices. The experiments are performed on a 200 MHz Pentium Pro running Linux.

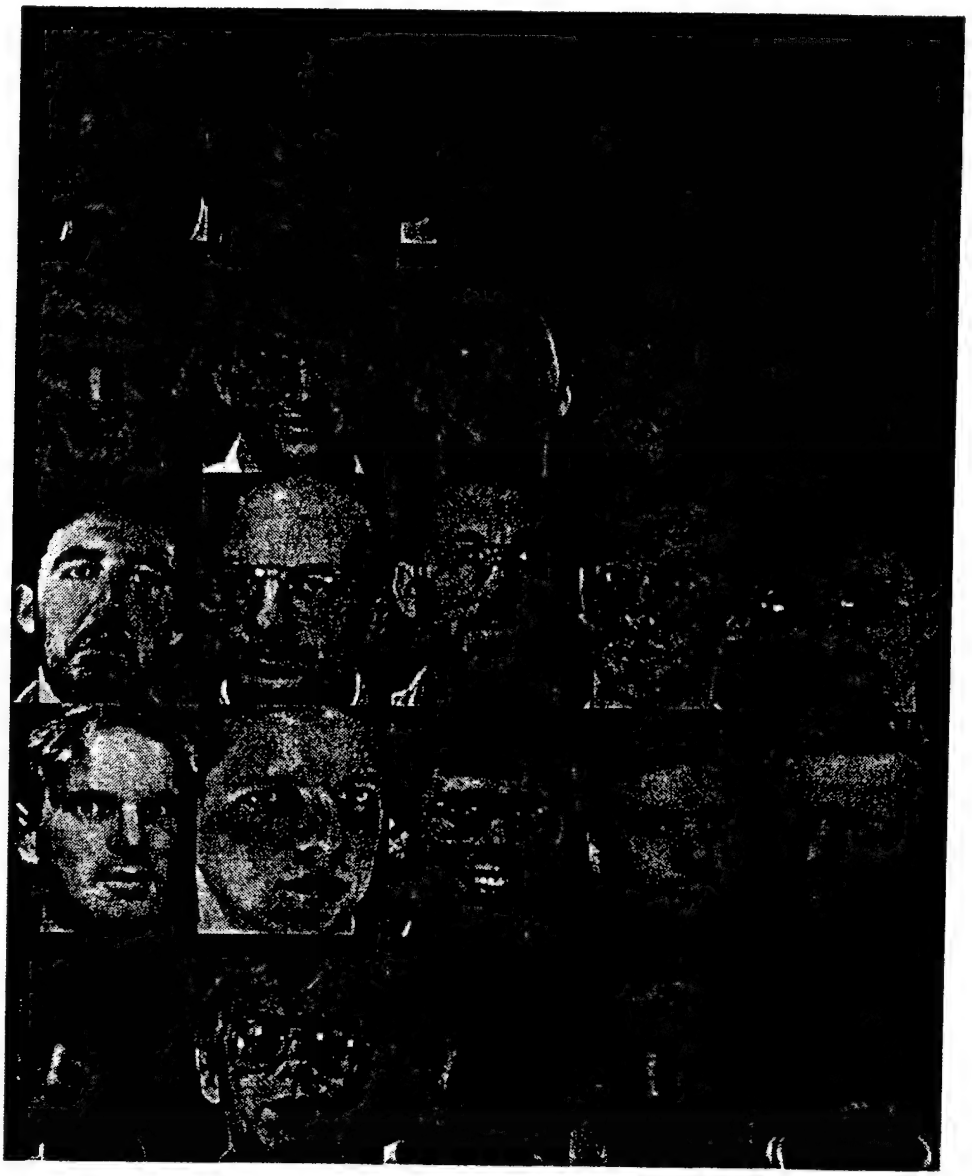


Figure 25: Shown are images of 25 faces from the set of test images which are used for the face recognition task using our matching scheme.

2.6 Summary and Conclusions

We present a novel representation technique for 3-D objects unifying both the viewer and model centered object representation approaches. The unified 3-D frequency-domain representation (called Volumetric/Iconic Spectral Signatures - V/ISS) encapsulates both the spatial structure of the object and a continuum of its views in the same data structure. We show that the frequency-domain representation of an object viewed from any direction can be directly extracted employing an extension of the Projection Slice theorem. Each view is a planar slice of the complete 3-D

V/ISS representation. Indexing into the V/ISS model is shown to be efficiently done using a transformation to a 4-D pose space of azimuth, elevation, swing (in plane image rotation) and scale. The actual matching is done by correlation techniques.

The application of the V/ISS representation is demonstrated for pose-invariant face recognition. Pose estimation and recognition experiments is carried out using a V/ISS model constructed from range data of a person and using gray level images to index into the model. The pose estimation errors are quite low at about 4.05° in azimuth, 5.63° in elevation, 2.68° in rotation and 0.0856 standard deviation in scale estimation. The standard deviation in scale is taken for the ratio of estimated size to true size. Thus it represents the standard deviation assuming a scale of 1.0. Face recognition experiments are also carried out on a large database of 40 subjects with face images in varying pose and illumination. Varying number of model images between 5 and 8 is used. Experimental results indicate recognition rates of 92.5% using 5 model images and goes up to 100% using 8 model images. This compares well with [37] who reported recognition rates of 87% and 95% using the same database with 5 training images. The eigenfaces approach [33] was able to achieve a 90% recognition rate [29] on this database. It also is comparable to the recognition rates of 96.2% reported in [29] again using 5 training images per person from the same database. These are highest reported recognition rates for the ORL database in the literature. The V/ISS model holds promise as a robust and reliable representation approach that inherits the merits of both the viewer and object centered approaches. We plan future investigations in using the V/ISS model for robust methods in generic object recognition.

References

[1] Z. Wang and J. Ben-Arie, "SVD and Log-Log Frequency Sampling with Gabor kernels for Invariant Pictorial Recognition," **Proceedings of 1997 IEEE International Conference Image Processing (ICIP'97)**, volume III, pages 162–165, Santa Barbara, CA, October 26-29 1997.

[2] D. Nandy and J. Ben-Arie, "Using the Fourier Slice Theorem for Representation of Object Views and Models with application to Face Recognition" **Proceedings of 1997 IEEE International Conference Image Processing (ICIP'97)**, volume III, pages 332–335, Santa Barbara, CA, October 26-29 1997.

[3] J. Ben-Arie and Z. Wang, "Pictorial Recognition using Affine Invariant Spectral Signatures," **Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)**, San Juan, Puerto Rico, June 17-19 1997.

[4] J. Ben-Arie and D. Nandy, "Representation of Objects in a Volumetric Frequency Domain with application to Face Recognition", **Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)**, San Juan, Puerto Rico, June 17-19 1997.

[5] J. Ben-Arie, Z. Wang, and K. R. Rao, "Affine invariant shape representation and recognition using gaussian kernels and multi-dimensional indexing," **Proceedings of 1996 IEEE International Conference on Speech, Acoustics and Signal Processing (ICASSP'96)**, volume 6, pages 3470–3473, Atlanta, Georgia, May 1996.

[6] J. Ben-Arie, Z. Wang, and K. R. Rao, "Iconic Recognition with Affine-Invariant Spectral Signatures," **Proceedings of 1996 IAPR/IEEE International Conference on Pattern Recognition (ICPR'96)**, volume 1, pages 672–676, Vienna, Austria, Aug. 1996.

[7] J. Ben-Arie, Z. Wang, and K. R. Rao. "Iconic Representation and Recognition using Affine-Invariant Spectral Signatures", **Proceedings of ARPA Image Understanding Workshop 1996**, pages 1277–1286, Palm Springs, CA, Feb. 1996.

[8] K. Arbter, W.E. Snyder, H. Burkhardt, and G. Hirzinger, "Application of Affine-Invariant Fourier Descriptors to Recognition of 3D Objects," **IEEE Trans. Pattern Analysis and Machine Intelligence**, Vol. 12, no. 7, pp. 452-459, July 1990.

[9] A. H. Barr. "Superquadrics and angle preserving transformations," **IEEE Computer Graphics and Application**, 1:11–23, 1981.

[10] D. J. Beymer, "Face recognition under varying pose," Technical Report A.I. Memo No. 1461, MIT Artificial Intelligence Laboratory, December 1993.

[11] J. Buhmann, J. Lange, C. von der Malsburg, J. C. Vorbruggen, and R. P. Wurtz, "Object recognition with Gabor functions in the dynamic link architecture: Parallel implementation on a transputer network," B. Kosko, editor, **Neural Networks for Signal Processing**, Pren. Hall, Englewd. Cliffs, NJ, 1992, pp. 121-159.

[12] R. Brunelli and T. Poggio, " Face recognition: Features versus templates," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 15(10):1042–1052, 1993.

[13] A. Califano and R. Mohan, "Multidimensional Indexing for Recognizing Visual Shapes," **IEEE Trans. on PAMI**, vol. 16, no. 4, April 1994, pp. 373-392.

[14] D. Casasent and D. Psaltis, "Position, rotation, and scale invariant optical correlation," **Applied Optics**, vol. 15, no. 7, July 1976, pp. 1795-1799.

[15] R. Cipolla and A. Blake, "Surface shape from the deformation of apparent contours," **Int. Journal of Computer Vision**, 9(2):83-112, 1992.

[16] I. Craw, D. Tock, and A. Bennet, "Finding face features," **Proceedings of European Conference on Computer Vision**, pages 92–96, 1992.

[17] J. G. Daugman, "Complete discrete 2D Gabor transform by neural networks for image analysis and compression," **IEEE Trans. on ASSP**, vol. 36, no. 7, July 1988, pp.1169-1179.

[18] D. DeMers and G. W. Cottrell, "Non-linear Dimensionality Reduction", in **Advances in Neural Information Processing Systems 5**, pages 580–587. S. J. Hanson, J. D. Cowan and C. L. Giles (editors), Morgan Kaufmann Publishers, San Mateo, CA., 1993.

[19] J. Flusser and T. Suk, "Affine Moment Invariants: A New Tool for Character Recognition," **Pattern Recognition Letters**, Vol. 15, pp. 433-436, Apr. 1994.

[20] J. Flusser and T. Suk, "Pattern Recogntion by Affine Moment Invariants," **Pattern Recognition**, Vol. 26, No. 1, pp. 167-174, 1993.

[21] R. M. Haralick and L. Shapiro, **Computer and Robotic Vision, Volume II**, chapter 18. Object Models and Matching. Addison-Wesley Publishing Company, Inc., 1993.

[22] R. Hecht-Nielsen, "Fast k-nn search for robust ATR object matching," **Proceedings of ARPA Image Understanding Workshop 1994**, Monterey, CA, Nov. 1994, pp. 889-894.

[23] M. K. Hu, "Visual Pattern recognition by Moment Invariants," in **Computer Methods in Image Analysis**, Eds. A. K. Agarwal R. O. Duda and A. Rosenfeld. IEEE Computer Society, LA, 1977.

[24] A. K. Jain, **Fundamentals of Digital Image Processing**, chapter 10. Image Reconstruction from Projections, pages 431–475. Prentice-Hall Inc., NJ, 1989.

[25] R. Jain, R. Kasturi, and B. G. Schunk, "Object Recognition," in **Machine Vision**, McGraw-Hill, Inc., NY, 1995.

[26] J. J. Koenderink and A. J. Van Doorn, "The Internal Representation of Solid Shape with Respect to Vision," **Biological Cybernetics**, 32:211–216, 1979.

[27] M. Lades, J. C. Vorbrüggen, J. Buchmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," **IEEE Transactions on Computers**, 42(3):300–311, 1993.

[28] J. Schwartz Y. Lamdan and H. Wolfson, "Geometric hashing : A general and efficient model-based recognition scheme," **Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition**, Tarpon Springs, MD, 1988, pp. 335-344.

[29] A. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face Recognition: A Convolutional Neural Network Approach," **IEEE Transactions on Neural Networks**, 8(1):98–113, 1997.

[30] C. Nastar and A. P. Pentland, "Matching and recognition using deformable intensity surfaces," In **Proceedings of 1995 IEEE International Symposium on Computer Vision**, Coral Gables, FL, May 1995.

[31] Olivetti and Oracle Research Laboratory. The ORL Database of Faces. Technical Report http://www.cam-orl.co.uk/facedatabase.html, 1994.

[32] A. P. Pentland, "Perceptual Organization and the Representation of Natural Form," **Artificial Intelligence**, 28:29–73, 1986.

[33] A. P. Pentland, B. Moghaddam, and T Starner, View-based and Modular Eigenspaces for Face Recognition," **Proceedings of 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)**, 1994.

[34] R. P. N. Rao and D. H. Ballard, "Object indexing using an iconic sparse distributed memory," **Proc. of Fifth Intl. Conf. on Comp. Vision**, Cambridge, MA, June 1995, pp. 24-31.

[35] D. Reisfeld and Y. Yeshurun, "Robust detection of facial features by generalized symmetry," **Proceedings of 1992 IEEE/IAPR International Conference on Pattern Recognition**, volume 1, pages 117–120, The Hague, The Netherlands, 1992.

[36] A. Rosenfeld and A. C. Kak, **Digital Image Processing**. Academic Press, NY, 1982.

[37] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," **Proceedings of 2nd IEEE Workshop on Applications of Computer Vision**, Sarasota, FL, December 1994.

[38] D. Terzopoulos and K. Waters, "Analysis of Facial Images using Physical and Anatomical Models," **Proceedings of 1996 IEEE International Conference on Computer Vision**, pages 727–732, Osaka, Japan, December 1990.

[39] Q.M. Tieng and W.W. Boles, "Wavelet-Based Affine Invariant Representation: A Tool for Recognizing Planar Objects in 3D Space," **IEEE Trans. on PAMI**, Vol. 19, No. 8, pp. 846-857, 1997.

[40] Seibert Michael and Waxman Allen M, "Adaptive 3-d object recognition from multiple views," **IEEE Trans. on PAMI**, vol. 14, no. 2, Feb. 1992.

[41] H. Wechsler, **Computational Vision**. Academic Press, New York, NY, 1990.

[42] J. Weng, N. Ahuja, and T. S. Huang, "Learning Recognition and Segmentation of 3-D Objects from 2-D Images," **Proceedings of 1993 IEEE International Conference on Computer Vision (ICCV'93)**, pages 121–128, 1993.

[43] X. Wu and B. Bhanu, "Target recognition using multi-scale Gabor filters," **Proceedings of ARPA Image Understanding Workshop 1994**, Monterey, CA, Nov. 1994, pp. 505-509.

[44] A. L. Yuille, P. W. Halliman, and D. S. Cohen, " Feature extraction from faces using deformable templates," **International Journal of Computer Vision**, 8(2):99–111, 1992.

[45] C. T. Zahn and R. Z. Roskies, "Fourier Descriptors for Plane Closed Curves," **IEEE Transactions on Computers**, C-21:269–281, 1972.